Sailboat and Kayak Detection Using Deep Learning Methods

Ziyuan Luo

A thesis submitted to the Auckland University of Technology
in partial fulfillment of the requirements for the degree of
Master of Computer and Information Sciences (MCIS)

2022
School of Engineering, Computer & Mathematical Sciences

Abstract

Visual object recognition is one of the most important and tough problems in computer vision. It targets various visual objects within realistic and real-time images. In depth, deep learning has become a powerful method to extract features directly from input data, which has made great progress in identifying visual objects. Recently, machine learning methods based on deep neural networks play a pivotal role in the field of visual object recognition. In order to identify ships in digital image, the nets need to be trained with a set of labelled images. So far, great progress has been made in visual object recognition based on deep learning, but developing relevant modules is a thorny job. Therefore, in this thesis, we propose a designated methodology based on search neural structure (NAS) for the recognition of visual objects by using our own published datasets to improve the results of sailboat detection. In addition, we conducted data collection for sailboat and kayak detections so as to find the best parameters based on basic model of YOLOv5. In this thesis, we also compare the net architectures and seek the best one. We test the proposed model and compare it with others.

Keywords: Ship detection, deep learning, CNN, NAS, YOLO, datasets

Table of Contents

List of Figures

List of Tables

List of Abbreviations

DETR	DEtection TRansformers
NMS	Non-Maximum Suppression
CNN	Convolutional Neural Networks
YOLO	You Only Look Once
ROI	Regions of Interest
NAS	Neural Architecture Search
HOG	Histogram of Oriented Gradients
C-HOG	Circle Histogram of Oriented Gradients
SVM	Support Vector Machine
RGT	Radial Gradient Transform
DPM	Deformable Parts Model
DCNN	Deep Convolutional Neural Network
R-CNN	Region-Based CNN
SSP	Spatial Pyramid Pool
ROI	Regions of Interest
FC	Fully Connected
RPN	Region Proposal Network
AI	Artificial Intelligence
ANN	Artificial Neural Network
DNN	Deep Neural Networks
CFAR	Constant False Alarm Rate
SSL	Self-Supervised Learning
NLDF	Non-Local Depth Feature
GBVS	Graph-Based Visual Saliency
FRFT	Fractional Fourier Transform
HOSC	High-Order Statistical Curve
ReLU	Rectified Linear Unit
BCE	Binary Cross Entropy
RFCN	Region-Based Fully Convolutional Network

Adam	Adaptive Moment Estimation
IoU	Intersection over Union
GIoU	Generalized Intersection over Union
CSPNet	Cross Stage Partial Network
FPN	Feature Pyramid Networks
FPS	Frames Per Second
AP	Average Precision
mAP	Mean Average Precision

Attestation of Authorship

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person (except where explicitly defined in the acknowledgments), nor material which to a substantial extent has been submitted for the award of any other degree or diploma of a university or other institution of higher learning.

Signature: Ziyuan Luo Date: 28 June 2022

Acknowledgment

During my master's study, my parents provided financial and spiritual support, thank them. In addition, my university, the Auckland University of Technology, my professors and my classmates have given me selfless help.

I would like to express my deepest gratitude to my two supervisors: Wei Qi Yan and Minh Nguyen. This thesis was completed under the guidance of my primary supervisor Dr. Yan. From the selection of the subject to the final completion, Dr. Yan and Dr. Nguyen has always given me careful guidance and support in academics, who allows me to successfully complete my studies.

Thanks to Ms. Kandy Corvette and Ms. Leeann Corvette from Screen 2 Script Limited (NZ), I have learned invaluable experience from them.

Thanks to my friend Ms. Jiayi Chen. Without her, I wouldn't be who I am today.

Ziyuan Luo

Auckland, New Zealand

June 2022

Chapter 1

Introduction

This chapter is composed of five parts. In the first part, we introduce the background and significance of ship detection based on deep learning methods, other parts contain the research questions, followed by the contributions, objectives, and structure of this thesis.

1.1 Background and Significance

In the field of maritime traffic management, ship detection is very important for maritime monitoring (Bi, Liu, & Gao, 2010). The existing ship detection methods are mainly based on continuous virtual police (CFAR). This approach is based on the transportation division of land and sea, which limits the speed required by fishing vessels (Kang, 2019).

As one of the most important applications of remote sensing, ship detection plays a pivotal role in commercial, fishery, transportation, and military applications. In particular, polarimetric synthetic aperture radar (OSPAR) is very sensitive and effective for detecting ships, because it provides excellent sensors, collects a large amount of structure and texture information, which can be effectively analyzed in any wild weather. Therefore, in recent years, the discovery of ships in POLSAR images has attracted much attention. Adaptive detection makes virtual police as one of the most popular ship detectors. However, its performance largely depends on the experience of local background noises and selected windows of digital images and videos (e.g., target window, protection window, and background window).

As we all know, heterogeneous chaos and interference objects usually lead to inaccurate estimation. In order to solve the inherent problems of virtual alarm sensors, an improved alarm detector was proposed. By improving the background evaluation method, the imbalance of interference is reduced. This can be solved by intercepting statistical data (Smith, 2000).

In recent decades, as one of the most important and difficult tasks in marine intelligent transportation, computer-aided marine ship detection has attracted extensive attention. As shown in Figure 1.1, from 2012 to 2021, the number of publications on marine matters has been increasing. The data in the figure comes from Google scholar advanced search. The three achievements of data acquisition, computing power, and algorithms have promoted applications of the advanced knowledge in the field of maritime management.

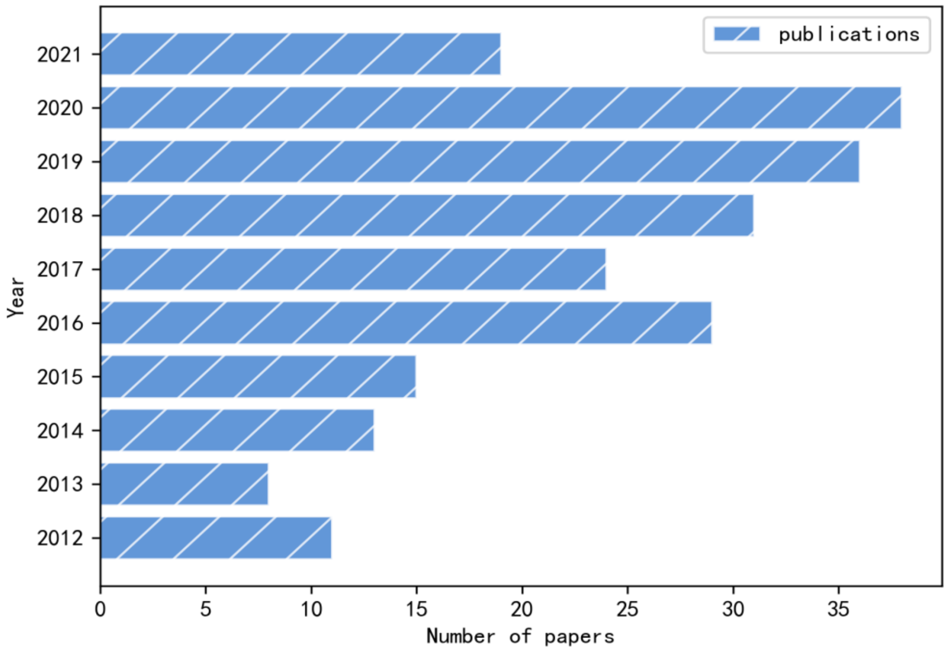


Figure 1.1: The number of publications in marine object detection

Deep learning has taken great step in recent decades which is becoming the most powerful technology in intelligent transportation. In numerous areas of maritime sector, including ship classification, target object selection, collision prevention, risk perception and anomaly detection, the methods for training deep learning models have been adopted. It is mainly employed for maritime surveillance and ship navigation. At present, the focuses are on the aspects of automated machine learning methods.

However, the methods cannot solve complex problems. Until now, most management processes in the shipping field are still dominated by human judgment, which is limited and inevitably error prone. Deep learning methods take use of data to obtain actionable information, and can provide more accurate classification. Based on this point of view, it is necessary to study the applications of deep learning methods in the marine and explore how computer vision can be explored or even surpass human ability in practice, especially in line with the rapid growth of visual object recognition in recent years.

With the latest progress of deep learning, more and more deep learning methods are being applied to smart ships (Khan et al., 2017). In 2020, Pan et al. proposed a deep learning-based RMA classification model that implemented navigation identification for intelligent ships and provided accurate navigation (Pan et al., 2019). The vision system, which takes use of computer vision to recognize ships and landmarks from digital images in navigation environment, has been evolved as an indispensable component of ship perception system (Chen, et al., 2017). As a result, an effective ship detection is critical for increasing the safety of maritime ships.

Because of fully automated feature extraction and representation capabilities, deep convolutional neural networks have achieved significant success in visual object detection via the end-to-end way (Bhandare, et al., 2016). Deep learning detectors are generally replying on big data, require little human participations, which are efficient and simple (Goodfellow, et al., 2016). The visual object detectors based on CNN can be grouped into two categories: One-step methods and two-step methods. The two-step methods have two phases: Positioning and classification.

Region-based CNNs (R-CNNs) for visual object detection outperform most existing methods (Girshick, et al., 2014). The two-stage detection algorithms encapsulate Faster R-CNN (Ren, et al., 2017), Mask R-CNN (He, et al., 2020), and R-FCN (Dai et al., 2016) which have greater accuracy but are very time-consuming and laborious. One-stage approaches carry out visual object detection directly while also classifying objects and conducting location regression. You Only Look Once (YOLO) is the first one-stage object detection method, which tackles the whole input image just once, reduces computing redundancy, and increases the detection speed (Redmon et al., 2016). Single Shot Detector (SSD) (Liu, et al., 2016), RetinaNet (Lin, et al., 2020), YOLOv3 (Redmon and Farhadi, 2018), YOLOv4 (Bochkovskiy, et al., 2020) and the newest YOLOv5 models convert pattern classification problems into regression problems.

Because of the end-to-end characteristic, the one-step methods are relatively fast and easy to be trained, which allows for real-time processing that is much suitable for mobile deployment. Deep learning approaches are increasingly being used in smart and fast ship detection. The focus is on enhancing detection accuracy, a majority of them, like Faster R-CNN (Li, et al., 2017) employ a two-stage detection architecture, which has made great enhancements to the original Faster R-CNN, such as adding negative mining and dense connection (Jiao, et al., 2018). There are various approaches for constructing a more complicated network in order to increase the resilience of particular challenges, such as dense tiny ships (Zhao, et al., 2018). Recently, more and more approaches have been implemented for high-speed processing of ship identification. Most of them, like the YOLO series, were constructed on a one-stage detection framework. Chang et al. firstly took use of YOLOv2 for ship recognition in SAR photos, which shortens the computational time (Chang, et al., 2019).

DCNN can automatically extract hierarchical elements from a large amount of training data, which has been successfully applied to pattern classification and visual object detection (Girshick, et al., 2014). It provides a fast and accurate detector. In other words, it is a region-based convolutional neural network (R-CNN). Spatial neural network is an artificial neural network composed of hierarchical nodes. The main difference is that we assume that the input is an image, which was employed to improve the execution time with the accuracy of image classification. Kang et al. proposed an improved Fast R-CNN algorithm based on CFAR algorithm.

In order to improve the detection of small ships, NRCan generated region proposals for the CFAR prediction windows more quickly (Kang, et al., 2017). It can also work well in homogeneous regions, but its detection performance in heterogeneous regions is poor due to the lack of CFAR. In order to further improve detection performance by using compression and inducement mechanisms, a new network architecture was proposed based on Faster R-CNN (Lin, et al., 2019). The test results show the effectiveness of recommended ship detectors, but many ships are missed.

On the other hand, the accuracy of bounding box detection generated by the proposed ship detector is low. Combining the expertise in polarization and object scattering mechanism, Chen et al developed a polarization-based object detection and classification system (Chen, et al., 2018). By using the idea of deep neural networks, a fast region-based convolutional neural network (R-CNN) method was proposed for vessel detection via high-resolution images from remote sensing (Zhang, et al., 2019). The bottleneck of NRCan lies in local proposals and selective search algorithm. The basic idea of accelerating NRCan is that local proposals must work with the characteristics of CDN well.

Therefore, we are use of feature maps extracted instead of selective search. The function mapping is sent to the regional proposal network (RPN), which takes use of the sliding windows and the CNN-based feature maps which sends out potential boundary boxes and accuracy predictions of these marked rectangles. The only difference between these two ways is that the local proposals only rely on Faster R-CNN. Therefore, it bypasses the selective search algorithm.

YOLO is based on Darknet framework (Redmon, 2016). Darknet is an open-source neural network programmed in C with the assistance of CUDA (Redmon, 2013). YOLO is claimed to be 100 × faster than Faster R-CNN. While the detection rate of YOLO is 45 frames per second, the Tiny YOLO model can be operated at 155 frames per second. This Tiny YOLO model only requires 516MB of GPU memory. The frame rates were achieved with the NVIDIA Titan X.

With the advent of deep learning along with the development of big data and GPU/FPGA, especially image processing methods have contributed greatly to computer vision, which detected ships from digital images as visual objects by using these advanced algorithms and facilities.

1.2 Research Questions

Ship detection has been employed in recent years. Retrieving the characteristics of a vessel and classifying is the basic procedure for detecting and recognizing a vessel. Therefore, the research questions of this thesis are as follows.

Question: *What methods can be applied to detect ships effectively*?

There are many factors that influence the detection outcomes, which make the detection results much difficult in the process of detecting a vessel. Based on these existing reasons, we have explored the following questions:

Question: *Which algorithm is right for ship detection? How can we make ship detection faster and better*?

The main idea of this thesis is to find a better way to improve the accuracy of detection results. In order to attain this, a plethora of advanced models and algorithms need to be explored and exploited.

Question: *Amongst the ships, which ones are chosen for this research project?*

This thesis was fulfilled in Auckland, New Zealand, so we selected the most culturally representative types of boats, which are sailboats and kayaks. There is still a research gap in the field of visual object recognition for these two types of ships.

1.3 Contribution

The contribution of this thesis lies in deep learning methods for ship detection. We experiment with real-time algorithms for ship detection. The experiments consist of five parts: 1) Data collection, 2) data augmentation, 3) defining the net structure and deployment, 4) model training, 5) algorithm evaluations and comparisons.

Moreover, in this thesis, we investigate which algorithm is the most suitable one for deep learning-based ship detection. Comparisons and analysis of various parameters are conducted within the same model.

In addition, the focus of this thesis is on deep learning models in line with CNNs because deep neural networks are developed mainly for pattern classification. At the end of this thesis, we will compare the results of our models by using the same dataset to prove the effectivity of the proposed algorithms.

1.4 Objectives of This Thesis

Firstly, we need to collect large-scale ship datasets. But we think it's too slow and may waste our time. In terms of datasets, our goal is to create a number of training datasets by segmenting the acquired images using reasonable data augmentation. Secondly, we avoid the influence of environment and plan to find out whether the proposed algorithm can recognize the position of a vessel.

1.5 Structure of This Thesis

The structure of this thesis is described as follows:

- In Chapter 2, we conduct a literature review and discuss the relevant issues of R-CNN. Meanwhile, we will deepen our understanding of attention models and Transformers in this thesis.
- In Chapter 3, we introduce our methodology. In this chapter, we will introduce the net design and deployment as well as experimental results and comparisons.
- In Chapter 4, we implement the proposed algorithm, collect experimental data, and demonstrate the research results. In addition, the limitations of the proposed methodology are depicted on details.
- In Chapter 5, we summarize and analyze the experimental results.
- We draw the conclusion and address our future work in Chapter 6.

Chapter 2
Literature Review

The focus of this thesis is on ship detection based on deep learning methods. In this chapter, we will introduce a plenty of conventional machine learning methods and the relevant knowledge of visual object detection, as well as datasets for model testing and performance evaluation.

2.1 Introduction

Auckland has a myriad of scenic ports, sailing as an important sport has been widely developed. Under the influence of sailing culture in this city, we have an idea to combine this modern sport with the in-depth study of artificial intelligence. The sailing object recognition from digital images already exists, but it still needs further research work for a better sailing object recognition.

Visual object recognition is one of the solidate foundations in the field of computer vision and also an important part of digital image and video processing. It automatically seeks the position of visual object in the given image through specific algorithms, and further determines the class or characteristics of the object. However, the diversity of visual objects, the irregularity of motions, the variety of visual angles and background and the styles of visual descriptions in digital videos can affect the final effect of visual object recognition. Therefore, the accurate detection and classification of objects are still an important research direction and has extremely high value both in academia and industry (Wang et al., 2019).

In computer vision, there are two well-known categories of methods for visual object detection that have been proven to be very effective. The first is R-CNN series algorithms based on region proposals (Girshick et al., 2014). This series includes R-CNN, Fast R-CNN, Faster R-CNN, and Mask R-CNN, which are all two-stage algorithms. The two-step algorithms have high accuracy but run very slowly. In order to resolve this problem, YOLO was introduced as single-stage algorithm. YOLO and its family greatly improve object detection speed with accuracy decreasing lightly. Since its birth, YOLO has received a few improvements and demonstrated betterment from YOLOv2 to YOLOv5. Owing to the significant improvement in detection speed and efficiency, YOLO and its family have been rapidly developed in recent years.

Therefore, in this thesis, we follow this series of algorithms and make our improvements based on YOLOv5 to balance speed and accuracy for sailboat detection.

In recent years, Transformers have been developed and mixed into deep learning mechanism based on a simple and powerful mechanism, which enables us to focus on the inputs. Detection Transformer (DETR) is the first application of Transformer in visual object detection (Carion, 2019). In Microsoft COCO dataset (Lin et al., 2014), the accuracy and speed of DETR are equivalent to the optimized Fast R-CNN (Xu et al., 2017). However, in terms of large object detection, DETR has a better performance than Fast R-CNN. In addition, unlike most existing object detection methods, DETR does not require a non-maximum suppression (NMS), which treats visual object detection as a prediction problem and thus achieves the model training in the end-to-end way. Our experimental results also show that the combination of Transformers and CNN models together has generated outstanding outcomes.

In this thesis, we combine convolutional neural networks (CNNs) with monitoring mechanism, so that we can concentrate on regions of interest (ROI) with significant potentials. However, designing a suitable model for visual object detection is a tough task which takes use of a lot of experiments to get continuous improvement. However, these processes do not always produce better results.

With the advent of neural architecture search (NAS), the research problem has been resolved (Elsken et al., 2017). By determining appropriate search and evaluation strategies, NAS search can design and evaluate network modules and ultimately obtain the best network model. By considering the shortcomings of facility identification, monitoring mechanism and solution automatic design, we incorporate them into this thesis to achieve the best efficiency. Our contributions are:

- Trough collecting data manually, we construct a realistic sailboat dataset, which allows us to evaluate the robustness of our model in real-time applications.
- In order to better focus on ROI region, we propose a CNN model that combines spatial attention mechanisms together.
- Faced with the diversity of data collected by various devices, we create the model by automating the search and design to make the model much more robust, in spite of tedious procedure of attention module design.

In summary, our proposed model improves the YOLO in terms of the attention mechanism which is validated at present based on our own datasets.

2.2 Generic Object Detection

Although our visual environment contains countless objects, the current focus of the research community is on identifying highly organized visual objects (e.g., vehicles, faces, bicycles, and aircraft) and jointly-related visual objects (e.g., people, cattle, and horses), rather than static scenes (e.g., sky, grass, and clouds).

As shown in Figure 2.1, visual object detection is grouped into two categories: Static object detection and dynamic object detection. The first group is to determine a match for a particular object (e.g., mountains, the Eiffel tower, etc.). The second one is to inspect specific objects (e.g., people, cars, bicycles, and dogs).

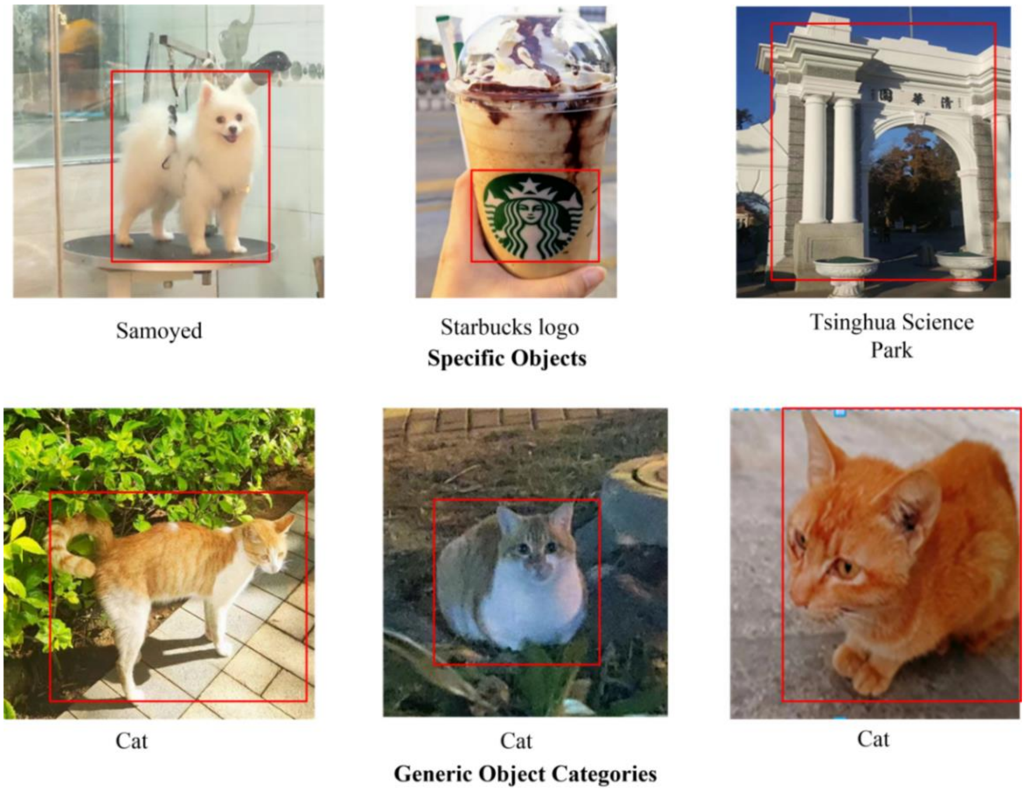


Figure 2.1: The types of visual object detection

Spatial position and size of a visual object can be represented by using bounding boxes, which means, accurately segmenting the pixels into content or closure boundaries. As far as we know, the bounding boxes are most broadly applied to visual object detection as stated in surveyed literature.

Many problems are closely related to the detection of visual objects. The purpose of object classification is to evaluate whether an object exists in a kind of image sets. This means that the specific objects need to be clearly classified into one or more classes with their locations.

The additional requirements of finding objects in images are more difficult than object classification. Object detection is a task in image processing, actually it belongs to pattern classification. Visual object detection is closely related to image segmentation, which specifies the semantic class for each pixel in the image.

Viola-Jones object detection framework was proposed in 2001 (Viola & Jones, 2001). The framework is based on AdaBoost algorithm (Ratsch et al., 2001), which detects human faces using Haar-like wavelet and integral graph. The object detection approach based on Haar + AdaBoost is also the first real-time framework. Prior to the emergence of deep learning, Viola-Jones detector became the industry standard for human face detection (Yang, et al., 2014).

Pertaining to histogram of oriented gradients (HOG), the histograms are calculated based on gradients rather than intensity of each pixel (Dalal & Triggs, 2005). It creates the feature vector by summing the gradient directions in histograms of local region in the given image. HOG features paired with SVM classifiers are incorporated into visual object detection, particularly the pedestrian detection (Wang, et al., 2008). Similar studies have been conducted to explore the attributes such as the invariants of HOG (Luo et al., 2015), which encapsulates spatial bins, the radial gradient transform (RGT) provides gradient invariance for visual feature descriptors.

DPM algorithm takes use of the improved HOG detector, SVM classifier, and sliding windows (Felzenszwalb et al., 2008), a component model of graphic structures is employed to solve the problem of visual object deformation. DPM is a component detection method that has a high resistance to deform visual objects. Currently, DPM has become the core algorithms for object classification, segmentation, pose estimation, etc. (Liu et al., 2016).

In specific applications, conventional machine learning-based visual object detection models, where image data is segmented into smaller blocks and expressed as vectors (Dolapci & Ozcan, 2021), may still retain an advantage. The feature vectors are created sequentially by adding components extracted from the colors and textures of the given images. A classification accuracy 99.62% was achieved by using the random forest method in Pascal VOC2012 dataset. On Apache spark, an average speedup 3.4 times was achieved while running each method on a 1 Master + 4 Worker clustering architecture. Figure 2.2 shows that this subject has been conducted in the past decades, meanwhile tremendous progress has been achieved.

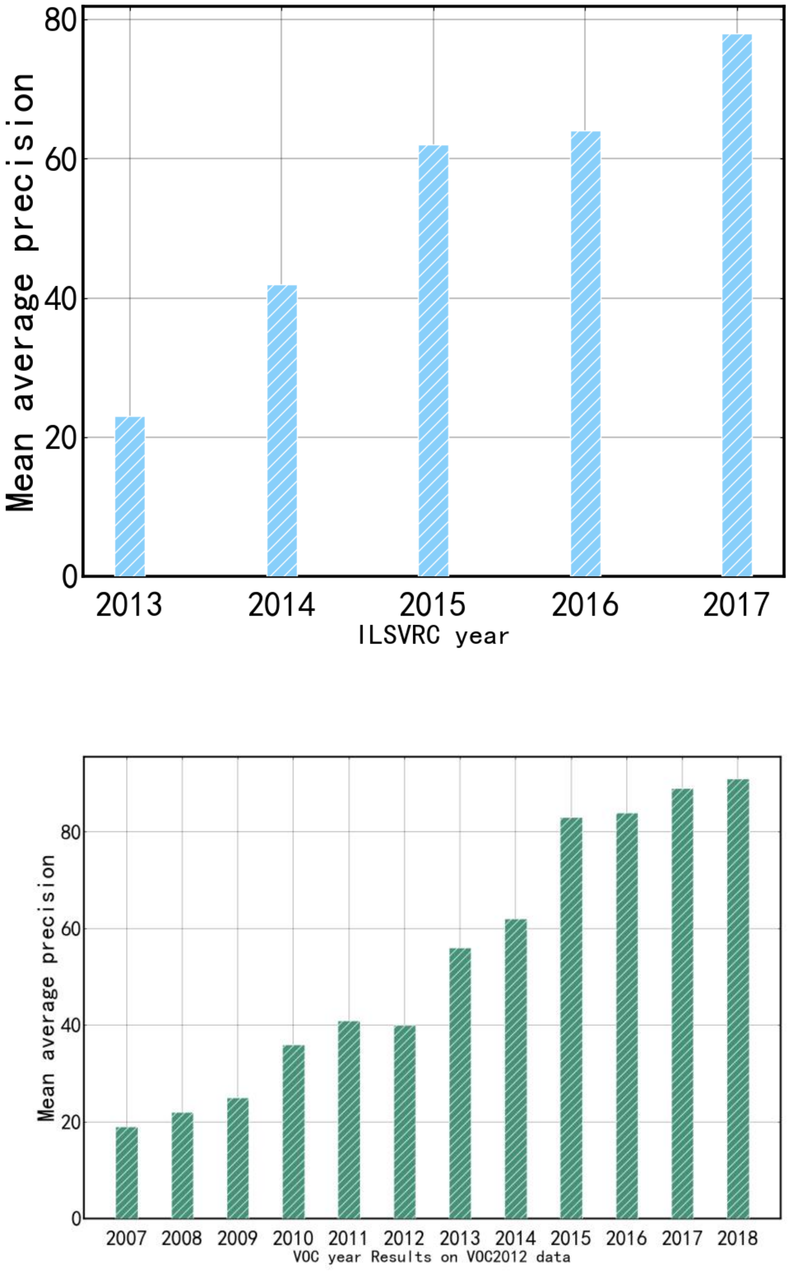


Figure 2.2: An overview of visual object detection algorithms

2.3 Object Detection Frameworks

The drastic shift from specified features to DCNN feature maps (Vedaldi, et al. 2009) demonstrates consistent improvement in visual object detection and recognition (Dai et al. 2016). In contrast, the fundamental "sliding window" (Felzenszwalb, et al. 2010) remains popular for searching locations, albeit the efforts to prevent exhaustive search (Uijlings et al. 2013). The number of windows, on the other hand, is vast and rises quadratically with the number of image pixels. Therefore, it is necessary to search across regions with numerous sizes and aspect ratios which will expand the search space. As a result, the development of efficient and effective object detection frameworks is critical to cut the computational cost.

In the proposed structures, region proposals are generated from the given images. CNN feature maps (Krizhevsky et al., 2012a) are extracted from those regions, then a classifier is employed to classify the proposal with a class label. DetectorNet (Szegedi et al., 2013), OverFeat (Sermanet et al., 2014), MultiBox (Erkhan et al., 2014), and R-CNN (Girshick et al., 2014) were independently and simultaneously proposed to detect visual objects through using CNN.

Girshick et al. was inspired by the end-to-end feature extraction algorithms and the success of selectively searching local proposals (Girshick et al., 2016), thus developing a R-CNN by combining AlexNet with the regional selective search method. Despite its ability to recognize visual objects with great accuracy, R-CNN has following limitations (Girshick, 2015):

- Model training is a multistage process with slow speed which is difficult to be optimized, because each stage must be independent on another.
- Model training for SVM classification and bounding box regressions is expensive in both storage and computing, because CNN feature maps must be extracted from visual objects specified in each image, which has a series of barriers due to large-scale dataset, especially in those very deep networks (Simonyan & Zisserman, 2015).
- The tests usually spend a long time, because the CNN feature maps are extracted

from the input visual objects in each test.

All of these disadvantages have driven further advancements, result in a variety of enhanced object detection frameworks, such as SPPNet, Fast R-CNN, Faster R-CNN, and so on.

The extraction of CNN feature maps is a major bottleneck in R-CNN, which are obtained from thousands of distorted regions of each image. As a result, He et al. (2014) introduced a spatial pyramid pool (SPP) into CNN architecture (Lazebnik et al., 2006). Because the convolutional layer allows arbitrary size of input images, the requirement for a fixed-size image is only determined by using fully connected (FC) layer, He et al. added a SPP layer before the FC layer to get a fixed size of output based on the FC layer.

With this SPPNet, R-CNN is able to achieve significant speedup without sacrificing detection accuracy, because R-CNN only needs to run the convolutional layers once over the entire test image to generate a fixed-size output. Although SPPNet speeds up R-CNN through multiple ways, it does not significantly improve the speed in the model training. The fine-tuning of SPPNet also limits the accuracy of deep neural networks.

In Fast R-CNN, the speed and accuracy of visual object detection are improved while eliminating the limitations of R-CNN and SPPNet (Girshick, 2015). The fixed-size R-CNN takes use of the idea of splitting the calculations between region proposals, adding an aggregation layer between the convolutional layers and the FC layer to extract the fixed-size feature maps. After merging ROI (regions of interest) and FC layer together, the branches are associated as a whole output eventually. The probabilities for predicting object class and class-specific bounding box regression are supplied eventually.

Compared to R-CNN/SPPNet, Fast R-CNN significantly improves the computing efficiency. Generally, it shows three times faster in training and 10 times faster in testing. It has a better search output, and a single training process updates all network layers without extra storage requirement.

Despite Fast R-CNN significantly accelerates visual object detection, it is still dependent on region proposals of visual objects, the calculation is a new bottleneck in Fast R-CNN. Recent research outcomes have shown that CNN has a unique ability to locate visual objects in convolutional layers (Hariharan et al. 2016), which makes it possible to replace selective search by using region proposals. A Faster R-CNN architecture provides an efficient and accurate Region Proposal Network (RPN) to prepare regional proposals, which is inserted between CNN and ROI pooling layers compared with Fast R-CNN to improve the efficiency of the algorithm.

2.3.1 Convolutional Neural Networks

Machine learning as a part of Artificial Intelligence (AI), has been utilized to represent complicated tasks in computer vision like visual object detection and scene understanding. Machine learning can provide effective solutions to a spate of challenging problems by substituting specific feature-extraction methods with deep learning algorithms. Machine learning is becoming increasingly feasible for computer vision with the generation of massive data nowadays.

Artificial neural network (ANN) is inspired by biological nerves, as a prominent artificial intelligence method. The idea is to simulate the signal transmission between neurons through input-output connections in computers. An artificial neuron takes multiple inputs with matrix multiplication and outputs a result with an activation function.

A neural network consists of numerous layers, each contains multiple neurons between the output and input. Hidden layers are those that exist between the output layer and input layer which are called visible layers. Each neuron in the hidden layer of a feedforward neural network takes information from the neurons in the preceding layers as well as provides next layer with an output based on activation function. A full ANN is made up of neurons in each layer that are linked together in layer-based way. Figure 2.3 depicts a basic feedforward multilayer ANN.

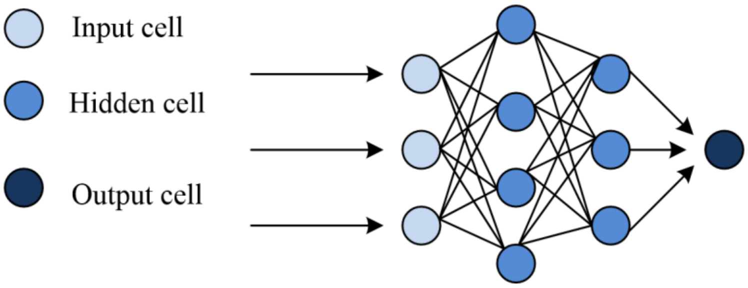


Figure 2.3: A simple feedforward multilayer artificial neural network

Deep learning has transformed a broad range of machine learning tasks, including visual object detection in digital image and video processing, as well as natural language processing and speech recognition. Given this incredibly quick progress, there are a plethora of recent surveys on deep learning (Wu et al. 2019). The research work examined deep learning approaches from various viewpoints (Zhou et al. 2018), natural language processing (Young et al. 2018), medical image analysis (Litjens et al. 2017), remote sensing (Zhu et al. 2017), and speech recognition (Zhang et al. 2018).

Deep learning is also named as deep neural networks or deep net. However, as its depth increases, the gradients may be vanished or exploded due to the backpropagation and differentiable properties of activation functions. In addition, in order to construct very deep networks, the number of samples in the training set ought to be quite huge. Furthermore, the computational costs for very deep networks may be very expensive, which require a long time to train the network model.

Deep learning is becoming practical as GPUs have become cheaper and more popular. As the growth of data accessibility, replacing conventional feature extraction approaches with deep neural networks will bring in a variety of usages. It is demonstrated that using the end-to-end way to initiate the layers of deep networks is much effective in feature extraction. As a result, the use of convolutional layers and pooling layers has become the mainstream for deep networks to extract feature maps.

Deep neural networks (DNNs), particularly convolutional neural networks (CNNs), are nonlinear mathematical models which are capable of extracting features for the representations of input images. CNN demonstrates the state-of-the-art performance on a variety of tasks in computer vision, including semantic segmentation, scene understanding, and object recognition, as well as the applications in remote sensing. In this section, we describe how to train a CNN model and produce correct results by using a majority of CNN advantages.

CNN is a sort of neural networks that have gained popularity because of its outstanding performance in computer vision and image processing. Commercial companies such as Microsoft, Google, NEC, AT&T and Facebook have established active groups to investigate novel CNN models. Currently, CNN-based models are being used by most of the frontrunners in computer vision (CV),　digital image and video processing.

Convolutional neural networks are employed in computer vision. When it comes to natural language processing, CNN models have shown excellent results in semantic parsing, text taxonomy, phrase modeling and classification, search query retrieval, text generation and prediction, as well as other classical natural language processing tasks. Besides, CNNs have been broadly employed to pattern classification. Medical image processing is one of the most popular applications of CNN, especially for cancer diagnosis using histological images. Spanhol et al. recently employed CNN to diagnose breast cancers and compared its results with a network trained on a dataset consisting of feature descriptors.

An input layer, multiple hidden layers, and an output layer contribute to a convolutional neural network. The hidden layer is made up of a sequence of convolutional layers. Convolutional layer is mathematically equivalent to a tensor product. Similar to the ordinary neural network, after the convolution calculation, the calculations will be spent on activation function. ReLU function is a popular activation function in convolutional layers of CNN models. A specific convolutional layer known as pooling layer is applied to downsampling the input of each convolutional layer. The CNN structure is accompanied with the fully connected layer at last after a sequence of convolutional layers and pooling layers.

DNN has demonstrated strong performance on datasets based on time series or grid-like architectures. However, there are additional issues that need to be addressed, such as the use of optimization methods to provide complete and reliable results for creating visualization models.

The primary difference between fully connected neural networks and CNN is that the CNN accepts the volume of 3D data or 2D images as the input, but fully connected networks treat the input as 1D feature vector.

2.3.2 Attention

Attention-guided context feature pyramid network (AC-FPN) is a new deep neural network, which not only expands the field of vision, but also captures discriminative semantics as well as locates precise positions through the attention mechanism (Cao et al., 2005). Squeeze-and-excitation networks (SENet) provide a multichannel monitoring mechanism in which the values between channels are defined by using two completely relevant layers (Jie et al., 2017). This will filter out unimportant data.

DETR is the first algorithm for visual object detection in Transformer model. It can better control large facilities, thus adversely affect small facilities. Recently, an oscillating Transformer has been proposed for visual object detection, which verifies the importance of observation data.

In addition, NAS has completed the design of monitoring mechanism. NAS modules and NAS modules are proposed to verify the weighted results by using the synchronous search. The model can search images in different locations by using the same network (Liu et al., 2020).

2.3.3 YOLO Family

YOLO creatively treated visual object detection as a single-stage regression task (Redmon et al., 2016). This model was developed in 2016, YOLOv2 was designed and inherited from Fast R-CNN and SSD algorithm, i.e., to classify patterns by using joint training activities (Redmon et al., 2017). YOLOv3 improved YOLOv2, deepened the network structure by using default model and multiscale structure of FPN (Redmon & Farhadi, 2018). YOLOv4 has developed an efficient and reliable model that allows anyone to use ultra-fast and accurate object detectors (Bochkovskiy et al., 2004). A series of deep learning methods for visual object detection have been verified. YOLO is a convolutional neural network that predicts the locations of bounding boxes and the class of visual objects at the same time. YOLO selects the whole image as the input rather than randomly picking up the sliding windows or region proposals to train the neural networks. The YOLO architecture is shown as Fig 2.4.

In 2018, YOLOv3 was developed (Redmon and Farhadi, 2018), which introduced several new and exciting concepts to deep network, including residual nets (He et al.,

2016) and feature maps (Lin et al, 2017). Whilst maintaining the speed of YOLO object detection, the accuracy of YOLO object detection is also increased. With the constantly advancement of deep learning methods, more and more methods have been proposed to improve visual object detection performance from various perspectives. Huang et al. presented an enhanced YOLOv3 model in 2020 to increase the accuracy of visual object detection (Huang et al., 2020).

In 2020, based on the original YOLOv3, integrated with excellent CNN optimization, including backbone, activation function, loss function, etc., YOLOv4 model for single-stage object detection was proposed (Bochkovskiy et al., 2020). Compared to YOLOv3, YOLOv4 model employs a robust data augmentation method, which includes Mosaic data augmentation and self-adversarial-training (SAT). In addition, the Mish activation function and CSPNet concept were developed based on the YOLOv3 backbone network to improve the backbone and feature extraction process.

Chen et al. (2021) proposed an improved YOLOv3 (ImYOLOv3) using the attention mechanism. However, there is no additional improvement of the speed with attention model incorporated into YOLOv3. Jie et al. (2021) presented *k*-means clustering algorithm and non-maximal suppression algorithm for YOLOv3 optimization to make it much suitable for ship detection, but the improved method is dependent on fine-tuning that was not a solution to improve accuracy adaptively in the field of ship detection. Compared to YOLOv4, YOLOv5 has faster speed and higher performance, which consists of a family of models, including YOLOv5s, YOLOv5m, YOLOv5l, YOLOv5x, and YOLOv5x+TTA.

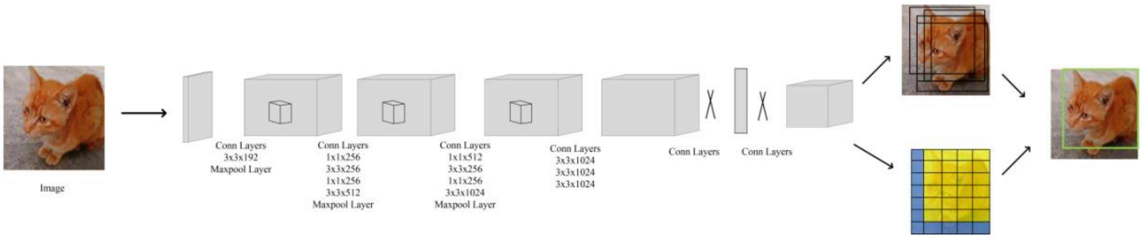


Figure 2.4: YOLO architecture

2.3.4 Neural Architecture Search

In 2016, Barker et al. proposed an intensive method to solve the design and optimization problem of attention model with CIFAR-10 dataset (Baker et al., 2017; Krizhevsky & Hinton, 2009). NAS can automatically create powerful neural networks based on multiple algorithms (Liu et al., 2018). The principle is that the candidate neural networks need to be established in the search space, which is a set containing all possible architectures, and then a specific optimization method will be used to search in the space for the best performing network structure. In the search process, the network structure is gradually optimized till the optimal subsystem is found. This method can save the design cost of deep neural networks.

This method is gradually employed to detect visual objects. Based on the results of the one-time NAS (Pham et al., 2018), DetNAS was proposed to conduct structural search (Chen et al., 2019). NAS and FPN are applied to solve the problem of how to automatically connect the neck and visual function so as to achieve a trade-off between accuracy and speed (Ghiasi et al., 2019). In addition, NAS and FPN take use of an improved search method, through using RNN as a controller with agents to accelerate the search. Auto FPN is related to the neck and head network. The neck is automatically integrated into the core functions, auto-head is use of the NAS search network to classify visual objects (Xu et al., 2019).

Overall, NAS has made significant progress in identifying and monitoring visual objects, but less progress in integration. In this thesis, we will continue improving the efficiency of sailing ship detection.

2.3.5 Transformer

Transformer is the approach that has gained popularity after the release of NLP and BERT model in 2018 (Devlin et al., 2018). The BERT model is pre-trained in a large text dataset, the focus of BERT is on diverse NLP tasks (Zhou & Tao, 2020). Transformer with multi-head method, on the other hand, is utilized for modeling after 2018 (Zhang et al., 2019). In the field of computer vision, there are two types of Transformer backbones. The DETR backbone is a typical approach for implementing Transformer as a pipeline with CNN in the object detection (Carion et al., 2020); Vision Transformer is implemented with Transformer pipeline with CNN not involved in its backbone structure (Dosovitskiy et al., 2021).

Attention is the base of Transformer (Vaswani et al., 2017). The Facebook team integrated YOLO with Transformer. The encoder-decoder structure is adopted between the backbone and the fully connected layers. There are also many usages of Transformer, for example, the original method can be improved by using Transformer (Zhang & Lin, 2013).

Seq2Seq approaches based on encoder-decoder scheme are useful for a variety of applications, including machine translation, text summary, and question answering. However, the contextual representations of encoders are ambiguous while dealing with far-reaching dependencies.

In order to solve these problems, Vaswani et al. proposed a unique architecture, called Transformer, which completely relies on self-attention to convert from its input to output without the sequence structure like RNN. The system structure of Transformer decoder is composed of encoder-decoder attention, self-attention and feedforward network. Without encoder-decoder attention, the decoder unit is actually the same.

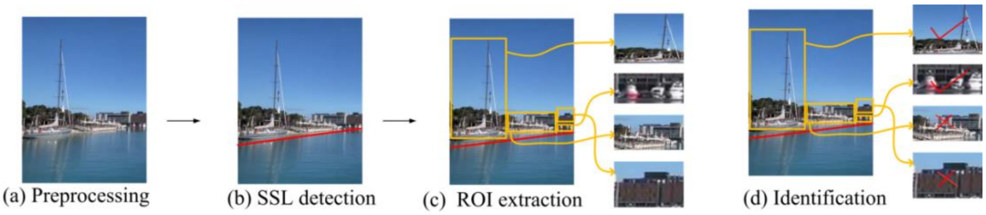
Based on the latest NLP operations, the base BERT model consists of 12 blocks (i.e., Transformer block), 768 hidden layers, and 12 self-attention heads; while the lager BERT consists of 24 blocks, 1024 hidden layers, and 16 self-attention heads. Although the performance of BERT model is very excellent, it still has a spate of shortcomings. Firstly, like other NLP models having transformers, the input sequences must be independent on input words and ignore all information about the location and dependencies between words. In other words, in order to predict markers, words and positions are embedded even if location information is a key aspect of NLP.

2.4 Conventional Ship Detection Methods

Conventionally, ship detection from remote sensing images depends on observation of scenes and theoretical methods based on statistics. The most popular ship detection system is based on constant alarm (i.e., constant virtual frequency), which fundamental assumption is that the background pixel intensity (non-ocean pixels) matches a known distribution, the gamma distribution. The concept of constant false alarm rate (CFAR) is calculated based on the self-construction to determine the detection threshold and obtain a group of false alarms, which can be implemented by using the algorithms that meet the characteristics of so-called CFAR. Therefore, the detectors based on constant false alarm rate are use of probabilistic interpretation to simulate the distribution of statistical pixels. However, these models are based on the processing of a single pixel, which leads to non-context decision-making that makes these detectors not suitable for high spatial resolution sensors on modern satellites, nor for images with dispersive objects.

Ship detection is thought as a subset of general object detection. Due to the unique properties of maritime environment, ship detection cannot fully conform to the general object detection paradigm. Deep learning has achieved outstanding performance in visual object detection. The visual detectors based on deep learning can be directly employed for ship detection; however, it is worth considering integrating the features of maritime surroundings. For clarity, in this thesis, we apply graphics to depict the workflow of ship detection, which includes both conventional and deep learning approaches.

Visual object detection takes use of sliding windows having various sizes on the input image to search for potential locations. Detecting the position of ships in the given image can efficiently reduce the search range of candidate locations and prevent interference. Furthermore, SSL (i.e., self-supervised learning) is employed as a significant indicator at the stages of ROI extraction and ship identification. The workflow of ship detection is shown in Figure 2.5. Prepossessing, SSL-based detection, ROI extraction, and identification are the four processes in traditional ship detection.



(c) ROI extraction

Figure 2.5: The workflow of conventional ship detection method in marine environment

2.4.1 Preprocessing

The initial stage in ship detection is to eliminate or limit negative effects. To filter noises, Tang employed Gaussian smoothing (Tang, et al., 2013). To reduce background clutter, Li employed morphological reconstruction based on opening-and-closing operation (Li, et al., 2019). Lu, et al. (2006) utilized a median filter to remove noises. Sun suppressed the background using a wavelet transformation (Sun, et al., 2005). To remove the change in the background, Bouma evaluated the background intensity and subtracted it from the input images (Bouma et al., 2008). It is difficult to totally remove environmental noises during preprocessing.

The image dataset is with tagged files. The dataset is composed of 3D images with a raw file, which contains location information. In order to use the dataset, the first step is to convert the MDB/raw file into a .jpg/.txt file, which stores the image information and the bounding box information corresponding to the image.

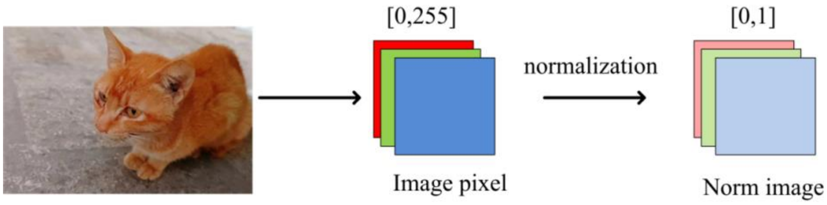


Figure 2.6: Image normalization process

Deep neural networks often have relatively moderate weights. The input of CNN models typically includes pixel intensity in the range [0, 255], which is usually integer. Zero-centering and normalization of the training dataset could be applied to improve network performance. Zero-centering could be accomplished by subtracting the mean of each image from training dataset, normalization could be achieved by dividing the pixel intensities of each input image by using standard deviation so that the pixel intensity of the training set is mapped between [0, 1].

Ship prescreening is one of the most important procedures in ship detection. The prescreening and discriminating steps are frequently combined to produce a label of single object detection. It is usually attained for efficiency, as SAR images can be quite huge, with the size up to 52400×37200 pixels, depending on image resolution.

After the SAR image has been appropriately referenced, the following step is to produce potential tags. The visual objects are recognized by applying a threshold to the image, which segments it into ships and non-ships. Due to the fundamental nature of these binary classification detectors, which need a trade-off between detection precision and false alarm rate, selecting a single threshold that delivers both high detection precision and low false alarm rate is sometimes impossible. Hence, SAR ship detection works for prescreening and discriminating stages. Image processing and statistical approaches based on SAR images are generally used in prescreening to identify regions that are likely to contain ships. Discrimination then utilizes the visual data as input to distinguish ships from ocean clutter and SAR artifacts to identify false alarms. Prescreening is grouped into three major categories: Global, local, and others, which will be addressed on details later.

The total accuracy of ship detection is heavily influenced by prescreening. Setting excessively strict global or local adaptive thresholding algorithm may fail to detect all probable ships in a given SAR image. Compared with local adaptive systems, global methods generate a larger number of false alarms at lower thresholds. The choice of CFAR is greatly determined by the images to be processed. In most instances, a single CFAR approach is sufficient for a single SAR dataset, but the criteria become difficult to achieve if the CFAR method is applied to a SAR dataset with various sensors, polarizations, and resolutions. As detailed in the following chapters, the simplest CFAR approach is adopted in this thesis, with an emphasis on decreasing false alarms.

2.4.2 SSL-Based Ship Detection

In marine circumstances, SSL serves as the border extractor between the sky and the water, which makes it easier to locate the ocean region and narrow down the search region. Figure 2.7 shows the process of reducing the candidate search regions after the SSL is conducted. Ships typically show around the edges in long-range object detection. The SSL-based approaches include additional processes such as ROI extraction. Prasad, et al.

(2017) and Lipschutz, et al. (2013) have summarized SSL-based edge detection approaches.

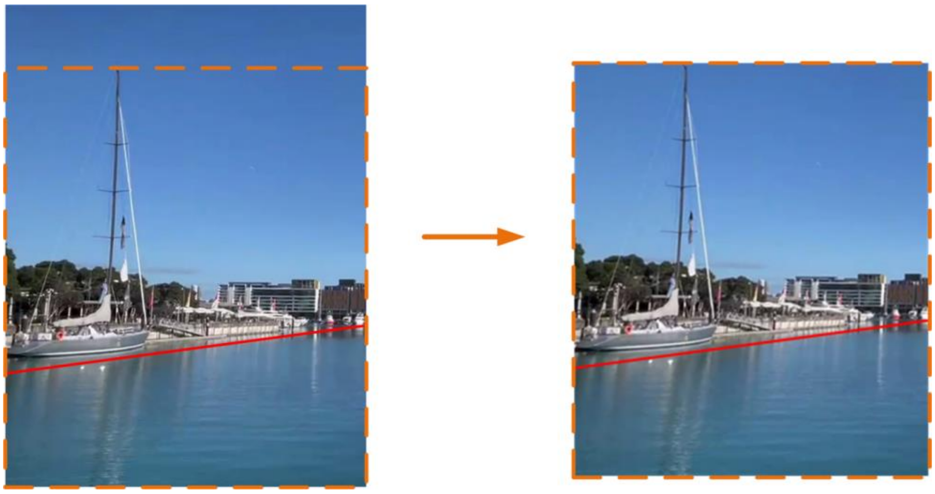


Figure 2.7: The reduced candidate regions for ship search

In transformation-based methods, in order to create an edge map, we firstly extract all edges in given images. Hough transform (Chen, et al., 2017) or Radon transform is then applied to the image so as to detect all edges. Finally, the SSL-based result will be generated and assessed based on a set of criteria. Tang, et al. (2013), for example, thoroughly evaluated the prospective sites by using gradient decent, gradient direction, and the average grayscale difference between the sea and the sky.

Visual object detection based on SSL is simple, but image edge extraction is required. Therefore, if the contrast of images is low or the sea-sky boundary is blurred, it is difficult to obtain a sharp edge image. In this case, the edge detector only responds to local changes that are easily perturbed by noises, which leads to the failure of edge detection and even the final object detection. Furthermore, uneven sea surface brightness can also lead to erroneous results in edge extraction, which in turn affects visual object detection.

In general, various regions of the sea or sky images may change slowly, but near the sea–sky border, the changes are drastic. Thus, the ship's position can be determined from changes of the given image. However, it is tough to categorize distinct regional description and identify the advantages and disadvantages.

This approach represents an image with various edges and segmented regions based on the differences of visual features. It takes global visual properties and local noises into account. To acquire a precise position of the ships, semantic segmentation can identify the pixels and produce a more detailed segmentation of regions (ocean, sky, and mixed region). To semantically segment objects from the given image, Yang utilized Gaussian mixture model (GMM) (Yang et al., 2019) while Jeong employed a neural network (Jeong, et al., 2018). A segmented image may be utilized to determine the sky and ocean regions as well as the location of the ships. This approach is typically complicated and requires labeled samples for model training. Semantic segmentation may not satisfy the real-time requirement of tiny or small devices with low computer resources due to the high intensive and amount of computing.

2.4.3 ROI Extraction

Followed the extraction of object location, the next step is to locate the region having candidate ships, i.e., the ROI. Thus, standard threshold approaches are utilized. Saliency detection is applied to extract ROIs from low-contrast images, such as thermal infrared images, and frequency-domain approaches are suitable for these images.

SSL-based detector is a useful way to determine ROI. Reasonable usage of SSL-based approach can significantly narrow down the scope of a region search. Ship detection is generally SSL-based, especially for long-range object detection. The possible ROIs are found around the object and generally are integrated with additional visual data to determine the correct ROI. Tang employed gradients and shapes (Tang et al.,2013), Lu adopted the distance rule (Lu, et al., 2006), while Chen utilized grayscale intensity (Chen et al., 2017). In order to extract ROIs, a fixed-height search region was set around the visual object (Fefilatyev et al., 2012) or segment the objects exactly near the target region (Shan et al., 2019). The placement mistakes, on the other hand, will affect the accuracy of ROI extraction. Therefore, the premise of using SSL-based method is to obtain the accurate position of the visual object.

Typically, there is a noticeable variation in intensity between ships and the background. A threshold can be employed to distinguish between ships and others. A conventional way for calculating ROIs is to use thresholds. In order to identify the right threshold, a variety of effective ways are identified. Wang investigated image complexity (Wang et al., 2010). Singh determined the necessary threshold by using language quantifiers (Singh et al., 2017). Bouma proposed hysteresis thresholding to evaluate the strength of the background and find the target (Bouma et al., 2008). When the contrast between a ship and the backdrop is minimal, particularly in thermal infrared images, it will be difficult to determine a suitable threshold. Furthermore, marine debris might have an impact on the threshold-based methods.

The contrast of images may be low and the intensity distribution might not be uniform. In order to identify ships efficiently, visual saliency detection may be utilized to determine saliency zones based on global information rather than only local information. Visual saliency detection methods based on visual data usually adopt a bottom-up approach to reduce the interference caused by local noise. These methods compare the color, brightness, edge, and other properties to determine the differences between the target region and its surrounding pixels.

To identify salient objects, a variety of methods are utilized. Li made advantage of intensity and contrast characteristics (Li et al., 2019). For low-contrast infrared images, Liu presented an enhanced non-local depth feature (NLDF) (Liu et al., 2019). Mumtaz employed a graph-based visual saliency system (GBVS) (Mumtaz et al., 2016). Lin proposed a number of visual elements, including gradient texture, brightness, and color aspects, to create a striking image (Lin C et al., 2020). If there is a vast of regions with a lot of marine debris, the saliency regions will be misidentified. As a result, reducing marine debris is a critical problem in ship detection.

In frequency domain, there are frequently noticeable discrepancies between the sea surface and a ship. Setting an appropriate threshold allows us to distinguish the saliency between the backdrop and the target. Fourier transform and wavelet transform are two extensively used frequency domain algorithms. Zhou proposed the fractional Fourier transform (FRFT) in conjunction with high-order statistical filtering (Zhou, et al., 2018). A high-order statistical curve (HOSC) isolates the target from the sea clutter. Sun employed the wavelet transform (Sun et al., 2005). If the sea congestion is dense, it is difficult to distinguish the objects between the backdrop and a ship by using only spectrum characteristics in frequency domain; however, if these approaches are coupled with morphological operations, the results may be generally better. Zhou has taken the characteristics into consideration to produce much accurate findings.

2.4.4 Object Identification

False alarms, such as sea clutter and islands, may be appeared in the ROIs. This outcomes of ship detection need to eliminate false alarms and select the appropriate candidate ships. The classification results are classified into two classes: Ship and non-ship. This kind of ship classifications were not paid much attention in the past.

Ships have defined proportions, lengths, and widths, hence, ships have prior information. Tang employed the aspect ratio, contrast ratio, duty ratio, and other factors to recognize ships from digital images. Ozertem eliminated false alarms based on prior knowledge related to the size of a ship and the size of the edges extracted based on the SSL (Özertem et al., 2016). Because a ship is generally close to an object shape and has the brightest grayscale intensity in an infrared image, the distance and grayscale intensity were coupled to assess whether a target is a ship or not (Lu et al., 2006). However, the robustness of ship detection based on past information is low, it is quickly disrupted by ocean waves, clouds, and islands, resulting in false alarms.

To generate more robust findings, feature vectors are employed to describe ships as samples. To characterize ships, Xu developed a rotation-invariant descriptor, a circle histogram of oriented gradients (C-HOG) (Xu et al., 2017). This description can feature infrared ships with various rotation angles. Finally, as a classifier, support vector machine (SVM) was employed for ship detection. Lin employed three types of features to characterize a ship: Size, form, and texture; a 10-dimensional feature vector was created to represent a ship; SVM was applied for offline training to efficiently reduce false alarms (Lin et al., 2019). This approach necessitates the selection of appropriate hand-crafted features, the classifier's training is dependent on the training samples.

2.5 Major Deep Learning Architectures for Ship Detection

Deep learning offers great opportunities in many fields, including biology and physics, not only in computer vision. At present, most modern object detection algorithms take use of deep learning network as the main basis to obtain features from digital images so as to complete classification and positioning tasks (Jiao et al., 2019). At present, ship detection based on deep learning usually adopts CNN-based methods. A slew of deep learning frameworks has provided basic components associated with APIs, which makes ship detection easier.

2.5.1 Single-Stage and Two-Stage Detectors

The conventional object detection methods based on artificial features still have room for improvement. Convolutional neural networks (CNN) can extract much detailed semantic information from images and produce stronger feature maps, AlexNet succeeded in the ImageNet competition (Krizhevsky et al., 2012). Since then, CNN has been applied to identify visual objects. Girshick launched R-CNN in 2014 and harnessed it for visual object recognition for the first time (Girshick et al., 2014). The current focus is on identifying visual objects through deep learning.

The object detection methods based on deep learning are split into two groups: (a) Two-stage detectors, e.g., R-CNN, Fast R-CNN (Girshick, 2015), and Faster R-CNN (Ren et al., 2016); (b) single-stage detectors, mainly YOLO (Redmon et al., 2017), YOLOv3 (Redmon et al., 2018), SSD (Liu et al., 2016) and DSD (Fu et al., 2017). The single-stage detector predicts the object class after feature extraction and returns the positions. Figure 2.8 shows the two-stage and single-stage detectors.

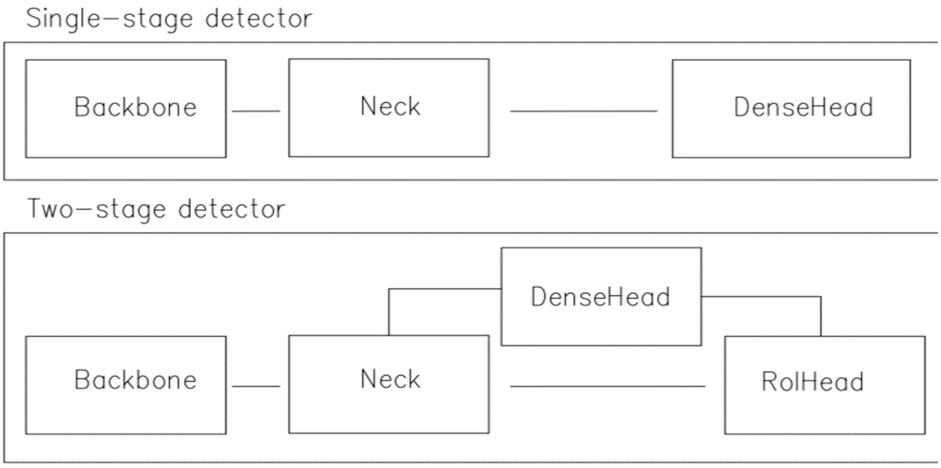


Figure 2.8: Two-stage detectors & single-stage detectors

2.5.2 Deep Learning in Ship Detection

In recent years, more and more academics have embraced deep learning for ship detection owing to the exceptional performance of CNNs in the field of visual object detection. The single-stage detectors have been widely employed in the research projects. The detection speed must be taken into account. For example, for the purpose of ship anti-collision warning, there are stringent criteria for real-time ship detection. Single-stage models often have faster response time and are able to tackle real-time work. Furthermore, many shipborne systems must take into account of the bulk and power consumption of computing equipment. Shipborne platforms often employ embedded devices for ship detection, as single-stage models require less processing time and consumes less power, which make them appropriate for deployment on embedded systems.

Two-stage detectors often offer higher detection accuracy but need much more calculations, thus lead to a low response speed. Shore-static systems often monitor ships at a port, where the moving speed of ships is sluggish and there is no need for severely limitation the computers and power usage. The two-stage models can be executed on a desktop computer.

Ships in an image have a diversity of scales because the viewing angle and distance are various. Ships come in a variety of sizes. The size of ships often changes, posing challenges for object detection networks based on pre-designed anchors (Shan et al., 2020). Furthermore, whilst identifying small marine ships from a distance, the ships usually occupy only a few pixels in the image, which is not conducive to feature extraction. Small maritime boats in an image may be overlooked, especially for a single-stage network.

In order to increase the effectiveness of detection networks in ship detection, the focus is now on two factors: Multiscale features and anchor settings. To increase the detection capabilities of deep nets for ships with various sizes, multiscale features or high-resolution feature maps are employed. In general, feature maps in a neural network include a pretty rich of structural and geometric information, which aid in object location regression (Oksuz et al., 2020). As a result, integrating multiple layers might yield complimentary information. Chen extracted feature maps with various sizes from a multilayer perceptron to improve the detecting capabilities of small maritime boats (Chen et al., 2018).

In order to extract feature maps from the detection network, high-resolution feature maps are employed. Hu established a feature pyramid network (FPN) using a scale transform module to deal with variations of ships induced by various imaging distances (Hu et al., 2019). Contextual information was utilized by blending high-level and low-level characteristics together. Shan utilized ResNet-50 as a backbone network and incorporated an FPN structure to overcome the problem of varying scales, which was particularly problematic for detecting tiny maritime boats. The feature maps generated by using three convolutional layers are fed into three region proposal networks (RPN), which improved models' discrimination ability with various sizes.

Anchors are a set of fixed-size initial candidate bounding boxes. The accuracy and speed of visual object detection and regression are affected by the size of an anchor. The variety of shape and size of ship, as well as differences in observation distance, provide a chance for anchor selection. The mainstream strategy is to examine the bounding boxes of ground truth in the training set and choose appropriate shapes and sizes for the anchor. K-means clustering is then harnessed to cluster all ground truth bounding boxes in the training dataset, the cluster centers are offered as starting point of anchors.

SSL is significant in ship detection that may effectively minimize the search area, which can remove interference regions. To the best of our knowledge, there is no application of the SSL in deep learning-based ship detection yet. The prospective candidate locations are filtered before importing an image to the detection network. Video streams are evaluated to identify significant spots of backdrops. Using SSL may effectively reduce candidate regions. The estimates are suitable for several lightweight SSL extraction techniques. Generating region proposals usually takes longer time in two-stage detectors. The SSL-based object detection can increase the efficiency of two-stage detectors and the detection accuracy.

2.5.3 Comparison of Conventional Methods and Deep Learning

Conventional ship detection methods do not require a huge number of training samples, a little amount of processing resources is highly interpretable. Conventional methods, on the other hand, are less adaptive to complicated settings and necessitate the selection of appropriate details of visual features.

Deep learning methods avoid selecting specific features in favor of automatically generating visual feature maps by using backbone networks to locate and detect visual objects. Deep learning algorithms are very adaptive to complex scenarios, particularly in marine environments where environmental frequently changes. A deep learning model, on the other hand, requires a high number of training samples to achieve better prediction outcomes; otherwise, overfitting may occur. There are currently few publicly available ship datasets. In terms of data scalability, there is still a significant difference between existing public ship datasets and generic object detection datasets.

2.6 Ship Detection Algorithms Based on CNN

The conventional object detection algorithm has a few shortcomings. Firstly, in order to obtain candidate regions, we need slide the windows, which leads to high computational complexity and long computational time, generates too many redundancies, and results in unnecessary over-computation. Secondly, the visual features lack adaptability to visual objects, the description of features may be inaccurate. Before Faster R-CNN algorithm was proposed, there were two popular algorithms based on convolutional neural networks, namely R-CNN algorithm and Fast R-CNN algorithm. Faster R-CNN algorithm has improved the previously existing algorithms, so before introducing Faster R-CNN algorithm, R-CNN algorithm and Fast R-CNN algorithm will be briefly introduced firstly.

2.6.1 R-CNN

Krizhevsky et al. showed us promising results of CNNs for the image classification challenge in 2012. Girshick et al. developed a model in 2013 that generalized these results for visual object detection. R-CNN refers to region-based CNN. There are multiple stages to R-CNN computations. Firstly, the regions of interest (ROI) are defined. Class-independent bounding boxes with a high probability contain an intriguing visual object. An approach called selective search is employed, though alternative region creation methods can be utilized instead. The visual features are extracted from each region by using a convolutional neural network. The bounding-box sub-image is warped to match the CNN's input size. The features extracted from the deep networks will be sent into support vector machine (SVM) for pattern classification.

R-CNN is significant since it is the first feasible solution for object detection based on CNN. There are three major issues for R-CNN. Firstly, the model training is split into multiple levels. Then, training the models requires a lot of computations and computer memory resources. The feature maps are generated from each region proposal and saved on disk for both SVM and region proposal regression. This will take hundreds of gigabytes of storage. Thirdly, the most important one, visual object detection is sluggish, it costs about a minute per image on a high-performance GPU. This is due to the CNN computing, the proposals for each visual object independently result in slow speed, even if the proposals are from the same image.

Because convolutional neural network is not accurate in feature extraction for visual objects, the region proposal method is proposed, which solves the problem of long operation time in conventional object detection algorithms. The region proposal method finds the candidate regions of a video frame that may be the target in the image by analyzing the color, texture, edge and other information. In this way, the number of candidate frames selected will be reduced, the size and position of the candidate frames will be relatively accurate, the quality of the obtained candidate regions is higher than that obtained ones by sliding the windows on video frames. Among them, selective search algorithm is one of the most widely used region proposal methods. The selective search is to extract candidate regions from input images according to their color, texture, and other features.

R-CNN algorithm is based on the combination of region proposal method and convolutional neural networks, instead of direct operations on extracting the video frames through sliding windows. As shown in Figure 2.9, R-CNN algorithm is to select 2,000 candidate regions by using selective search method. The sizes of the selected 2,000 candidate regions are not fixed, each region is input into CNN for model training and testing. Because the sizes of the input images are identical, after the convolution and pooling operations, it can be represented by using fixed-length feature vectors.

Finally, these feature vectors are classified by using SVM classifier, so as to find out whether the candidate region belongs to the visual object and what kind of class the visual object belongs to. Each candidate region has the probability that it corresponds to a class of objects. It is not necessary that each candidate region contains a visual object, as it may contain multiple objects in the same image. Therefore, each candidate region has different scores as confidence, we choose the maximum one so as to get the final bounding box correspondingly.

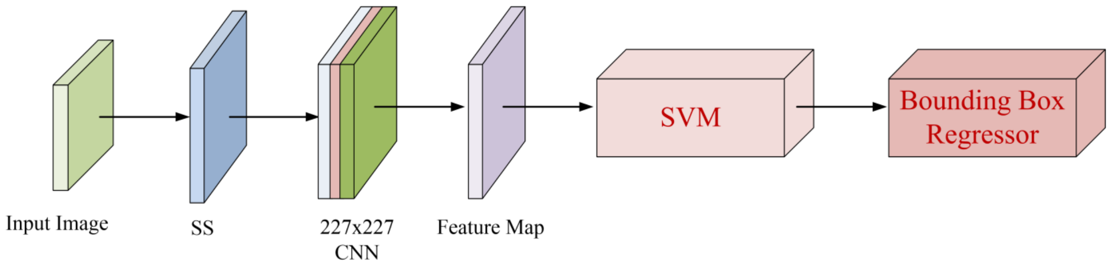


Figure 2.9: The block diagram of R-CNN algorithm

2.6.2 The SPP-Net Network

In SPP-Net network, pyramid pooling is applied to obtain the corresponding feature vectors of feature maps. Pyramid pooling is very important in accelerating the calculational speed and improving the detection accuracy of convolutional neural networks.

In the convolution operations of the convolutional neural networks, there is no requirement for the size of input images, but when it comes to classification and other operations in the fully connected layer, there is a strict requirement for the length of the input feature vectors. This means, the key problem is to make the feature maps of input images with the same length, so the spatial pyramid pooling (SPP) is proposed, whose structure is shown in Figure 2.10.

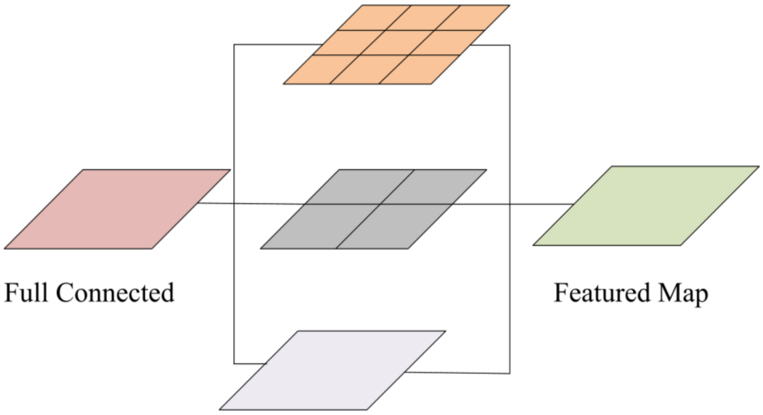


Figure 2.10: The pyramid structure diagram

SPP means splitting feature maps with various sizes into three scales (e.g., 4×4, 2×2, 1×1) so as to obtain feature vectors like 16×256, 4×256 and 1×256 dimensions. R-CNN algorithm needs to extract the feature maps using convolutional neural networks for each candidate region during model training and testing. However, the pyramid pooling only needs to extract the feature maps once, that is, to extract the feature maps of the whole image through convolutional neural networks, and finally to get the feature maps of each candidate region through pyramid pooling as well as the spatial correspondence between the feature maps and the original image. Because the whole algorithm only needs to be executed once through pyramid pooling, which does not forcibly redefine the size of the input images, the feature extraction will be much fast and accurate.

2.6.3 Fast R-CNN

The training and testing processes of R-CNN and SPP-Net networks are similar and relatively complex. In these two methods, firstly candidate regions are selected through selective networks, and then feature maps are extracted through convolutional neural networks, and finally are classified by using SVM classifier to get the results of visual object detection. However, the running time of R-CNN and the required resources are relatively large. Although it has been improved, compared with SPP-Net in the computational process, there are still 2,000 selected candidate regions. Therefore, the training time of SPP-Net has been shortened, the computer memory and data storage required in the training process are still very large.

The feature maps are extracted from the input image by using CNN in an end-to-end way, the CNN has been successfully applied to visual object detection in the input images. The idea of R-CNN is to still use the selective search method to generate candidate regions, then obtain the feature representation through convolutional neural networks, and utilize a classifier to get the classification results. However, the efficiency of this method is too low. It will take much time to use the generated candidate regions as the input of the convolutional neural network, and extract feature maps for visual object classification. In addition, the extracted candidate regions often have strong overlapping. If each region is input directly into the CNN for object classification, a lot of useless and redundant calculations will be carried out.

In order to save computing time, only one convolution operation is carried out to obtain the feature map, the feature maps are extracted from each candidate region based on feature maps. Fast R-CNN algorithm grows from Spatial Pyramid Pooling (SPP). By further improving pyramid pooling, Fast R-CNN puts forward Region of Interest (ROI).

An image is treated as the input, a series of region proposal networks are selected by using selective search, through a series of convolutional layers and pooling layers (e.g., VGG16) to obtain feature maps. Finally, the feature maps are obtained through pooling operations, the feature vectors with a fixed-length sub-sequence are treated as the input of fully connected layers corresponding to each candidate region. The fully connected layer is applied to visual object classification which has two outputs. One is related to the classification, i.e., the class of the visual object. The other output is related to the regression, i.e., the location of visual object is shown by the coordinates of the four corners of the bounding box of the corresponding region. The result is processed by non-maximum suppression to obtain class and location of the object.

The idea of ROI is to map the proposal to the position on the feature map. In this thesis, the size of pooling output is 7×7, so as to attain a fixed-size output for the region. Many experiments have proved that the image processing by pooling the region of interest is faster than the original R-CNN algorithm using pyramid pooling.

2.7 Datasets and Performance Evaluation

Throughout the history of visual object detection, datasets have played a pivotal role, not only include ground truth for evaluating and comparing the performance of algorithms, but also drive the relevant research work forward increasingly. In particular, deep learning approaches have brought enormous success to visual object detection in recent years, with a vast volume of annotated data which plays a decisive role in deep learning. The availability of enormous quantities of photographs on the Internet allows us to create comprehensive datasets, which encapsulate a tremendous richness and diversity of visual objects, enables exceptional performance in visual object detection and recognition.

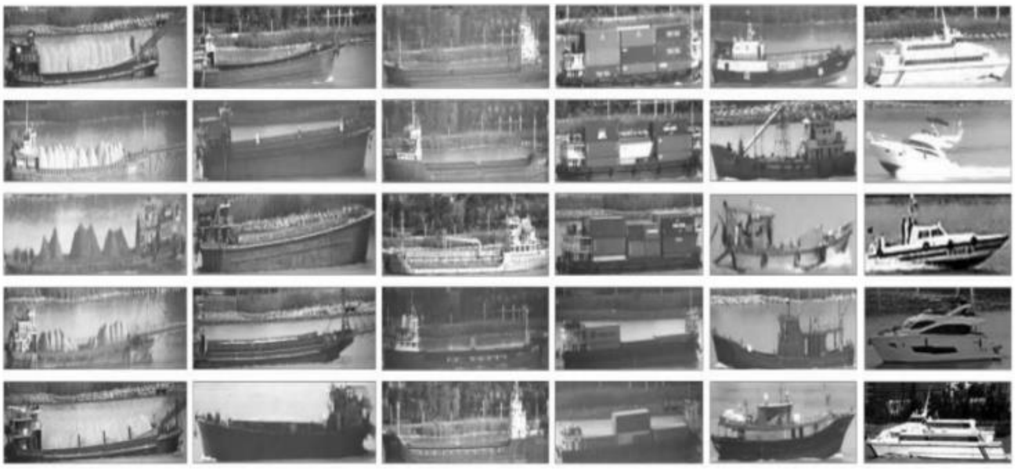


Figure 2.11: The public dataset of ships

There are four well-known datasets for general object detection: PASCAL VOC (Everingham et al., 2015), ImageNet (Deng et al., 2009), MS COCO (Lin et al., 2014), and Open Images (Kuznetsova et al., 2018). Figure 2.11 shows a part of what we have found from the ship datasets. Creating large-scale annotated datasets needs three major steps: Defining the class labels, collecting a wide range of photos to represent the selected classes on the Internet, annotating the acquired photographs, etc. (Kuznetsova et al., 2018). The four datasets served as the solid foundation for specific object detection. Each dataset includes a publicly available image collection, ground-truth annotation, standard assessment tools, etc.

PASCAL VOC (Everingham et al., 2015) is a multiyear project dedicated to the production and maintenance of a series of benchmark datasets for pattern classification and visual object detection. By starting with only four classes in 2005, the dataset has grown into 20 classes including all in our ordinary life. Since 2009, the number of photos has increased year after year, but all previous images have been retained to be able to compare the test results year after year. PASCAL VOC has increasing disadvantages owing to the availability of larger datasets such as ImageNet, MS COCO, and Open Images.

ILSVRC (Rusakovsky et al., 2015) is a derivative of ImageNet (Dan et al., 2009), which extends the standard algorithm of visual object recognition as well as evaluations. ImageNet1000, a subset of ImageNet images with 1,000 object classes and a total of 1.2 million images, has been modified to provide a standard test for the ILSVRC object classification.

The MS COCO database (Lin et al., 2014) contains complex scenes with visual objects in natural situations. In order to accurately estimate the detector, MS COCO object recognition comprises two object recognition processes which use bounding boxes as output or object instances as the segmented output.

Open Image Challenge Object Detection (OICOD) is a derivative of Open Images V4 (now V5 2019) (Kuznetsova et al., 2018) which is currently the largest publicly available object detection dataset. OICOD differs from earlier large object recognition datasets such as ILSVRC and MS COCO in terms of the annotation, as well as a significant increase in the number of classes, image, bounding box annotations, and instance segmentation mask annotations. In ILSVRC and MS COCO, all classes of instances in the dataset are fully annotated, while in OpenImages V4, a classifier was applied to each image, only labels with sufficiently high values are submitted for human validation. Therefore, OICOD only comments on positive labels confirmed by humans.

There are three criteria for assessing the effectiveness of visual object detection algorithms: Detection rate, accuracy, recall, and frames per second (FPS). The most popular metric is average precision (AP), which is derived from accuracy and recall. APs are usually rated by classes of visual objects. That is, it is calculated separately for each class of visual objects. To compare performance across all object classes, the mean AP (mAP) for all object classes is harnessed as the measure of performance (Everingham et al., 2015, Rusakovsky et al., 2015, Hoiem et al., 2012).

Chapter 3 Our Methods

The main content of this chapter is to clearly illustrate ship detection methods and articulate research methods which satisfy the objectives of this thesis. This chapter mainly covers the details of research methodology for ship detection using deep learning which will be clearly introduced with the confidence and imaginative use of the feature description methods.

3.1 Faster R-CNN

As shown in Figure 3.1, the framework of Faster R-CNN algorithm is split into four main parts, namely, feature extraction of the whole image using CNN, extraction of candidate regions using region proposal network (RPN), classification of candidate regions, bounding box regression, and non-maximum suppression. The framework of Faster R-CNN algorithm is firstly to obtain the feature maps of the whole image based on CNN, generate the candidate regions through RPN, and then use ROI Pooling to obtain the feature maps of each candidate region with a fixed-length feature vector. Finally, softmax is offered to classify with its result tackled by using non-maximum suppression (NMS) method to obtain the exact class of visual object.

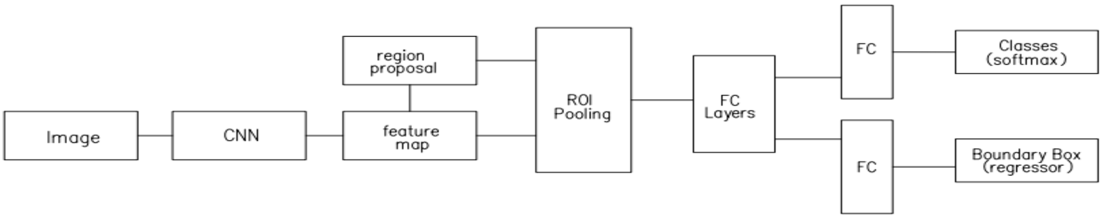


Figure 3.1: The framework of Faster R-CNN

3.1.1 Feature Extraction Network

In this chapter, we introduce open source deep neural networks, such as AlexNet, VGG, and other network structures, with the parameters trained by using ImageNet dataset. Because of the huge size of the ImageNet, the parameters in the network frameworks can be initialized well by using transfer learning method. By considering the trade-off between complexity and classification accuracy of the network, the feature extraction network selected in this thesis is VGG-16, due to the limited number of datasets, the weights trained in ImageNet are adopted as initial values for transfer learning. Our experiments show that transfer learning is very effective for the initialization of deep neural networks.

3.1.2 Region Proposal Network

Generally speaking, Region Proposal Networks (RPN) is an improvement of extracting candidate regions. Before its proposal, the traditional method for candidate region extraction is based on multiscale windows. However, candidate regions generated from these models often are overlapped. Pooling method is used for solving this problem later, but it will strongly affect the running speed, so we abandon the strategy in this thesis. Another relatively new method for candidate region extraction is selective search, but this algorithm will produce redundant calculations.

In order to reduce the calculations of redundancy, in Faster R-CNN, RPN is innovatively proposed to replace original selective search network. The pooling is set behind RPN to obtain the exact candidate region. The main structure of RPN is shown in Figure 3.2.

Firstly, the points on the feature maps are mapped to the original image with different sizes and shapes. Then the data is applied to train the region proposal network. According to the appearance of these bounding boxes and visual objects, whether they are target objects or not is identified finally.

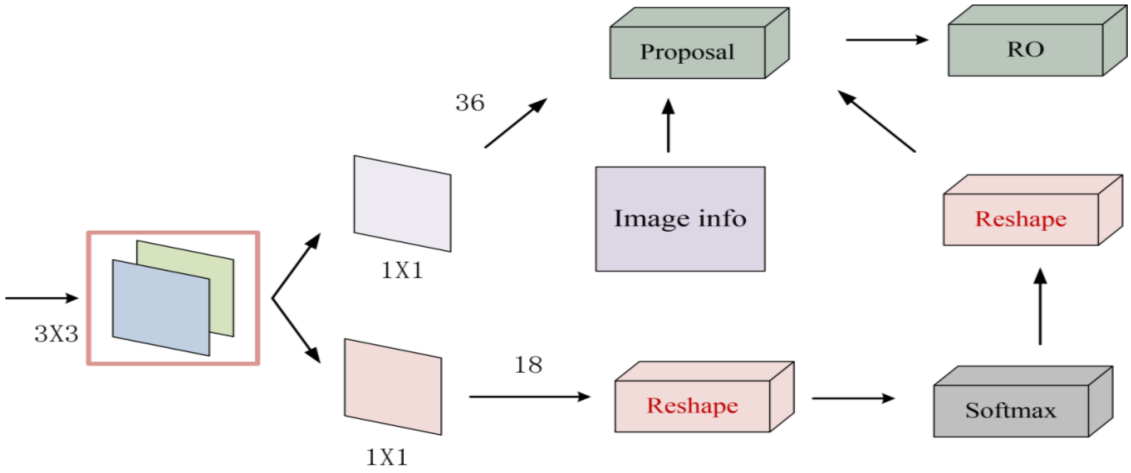


Figure 3.2: The suggested structure of regional proposal network

3.1.3 Classifications

Before classification, Faster R-CNN still takes use of the pooling method and outputs the fixed-length feature vectors corresponding to the FC layer-based classification. The classification is conducted with feature vectors obtained from the pooling operations, which are classified through the fully connected layer and softmax classification layer, then the model will output which class each region belongs to. Meanwhile, through bounding box regression again, the offset between each region and actual object position is obtained. This offset will be employed for subsequent regression to bring the bounding box of the detected visual object closer to the ground-truth position.

3.1.4 Non-Maximum Suppression

Non-maximum suppression (NMS) is applied in a few projects related to computer vision, such as edge extraction and visual object detection. After the classification, there will be a lot of candidate regions which will be identified as positive samples, that means, a number of candidate regions will be converged to visual object as the ground truth, there will be overlapping between them, the non-maximum suppression method is applied to solve this problem. The essence of non-maximum suppression is to select the maxima in the local range and suppress the elements that are not the exact maximum.

The steps of the non-maximum suppression algorithm are listed as follows: (1) Sort the scores of all the boxes and select the box with the highest score. (2) Traverse the remaining bounding boxes and set a threshold. If the intersection over union (IOU), is greater than this threshold, delete this bounding box. (3) From all the unprocessed bounding boxes, select the one with the highest score, and repeat the first two steps. After these three steps, we can basically achieve the goal.

3.2 Training Methods

There are two training methods for Faster R-CNN algorithm. In this thesis, we are use of the complicated training methods, the steps are listed as follows:

- **Step 1**: Firstly, we are use of the weights of CNN trained based on ImageNet dataset, and train a regional suggestion network independently by using transfer learning.

- **Step 2**: Train the Fast R-CNN network with the candidate regions generated in Step 1 as the input.
- **Step 3**: The region proposal network is retrained with the parameters of Fast R-CNN. But the parameters of convolutional layers shared by the region proposal network and Fast R-CNN are retained. The parameters related to the region proposal network are retrained.
- **Step 4**: Keep the convolutional layers shared by the region proposal network and Fast R-CNN network, fine-tune the parameters of those layers that are then subject to the Fast R-CNN, and finally implement the fast and accurate visual object detection.

3.3 Sailboat Detection Based on Automated Search Attention Mechanism

The algorithm that combines convolutional neural networks (CNN) with monitoring mechanism enables us to focus on regions of interest (ROI). However, it is problematic to design a suitable model for various tasks that require many experiments and continuous improvement. This process is arduous and does not always yield better results. With the emergence of neural architecture search (NAS), this problem has been solved. By determining the appropriate search strategy, search and evaluation space, the network module can be designed and evaluated, finally the best network model can be obtained. In this thesis, our main contributions are as follows:

- Based on actual data collection, we have established a reliable sailing ship detection dataset, through which we can evaluate the stability of our model in practical application.
- In order to better focus on the ROI of the models, we propose a CNN model that combines spatial attention mechanisms together.
- Considering the different data collected through different devices and the process of developing long focus modules, we create the model by automating the search and design, which makes it very robust.

In summary, our proposed model improves the YOLO family with the attention mechanisms which is validated at present based on our own datasets.

Attention model helps us to create a very accurate method for visual object detection by using NAS that will help us correctly design the best network for various datasets and solve the problem in visual object detection. Therefore, in this section, we will discuss the methods for sailboats detection.

3.3.1 Backbone

YOLOv5 is the last model of YOLO series. Although its performance is slightly weaker than YOLOv4, its service is quite long (Liu et al., 2021). This is the reason why we choose this network as the main supporting network in this thesis.

Pertaining to YOLOv5, the trunk, neck, and outlet are the same regardless of the version of YOLOv5s, YOLOv5m, YOLOv5l or YOLOv5x. The only differences between them are the depth and width of the model.

We provide a module as the first base layer. Its main function is to regularly extract pixels from high-resolution images and restore them to low-resolution images. We superimpose four adjacent image points, focus on the width and height of channel space, improve the recording domain of each point, and reduce the loss of original information. This module was designed to narrow down the calculations.

The third core layer, CSP network (i.e., cross stage partial), is composed of two core components: Bottleneck and CSP. The SPP module (i.e., spatial pyramid pooling) takes use of the largest pool, in turn, which is used to improve the perception of target objects.

The neck region allows a framework that provides information propagation based on R-CNN and FPN (i.e., feature pyramid networks) to accurately store spatial information, thus correctly define the pixels that create the mask. Later, the basic experimental model is established.

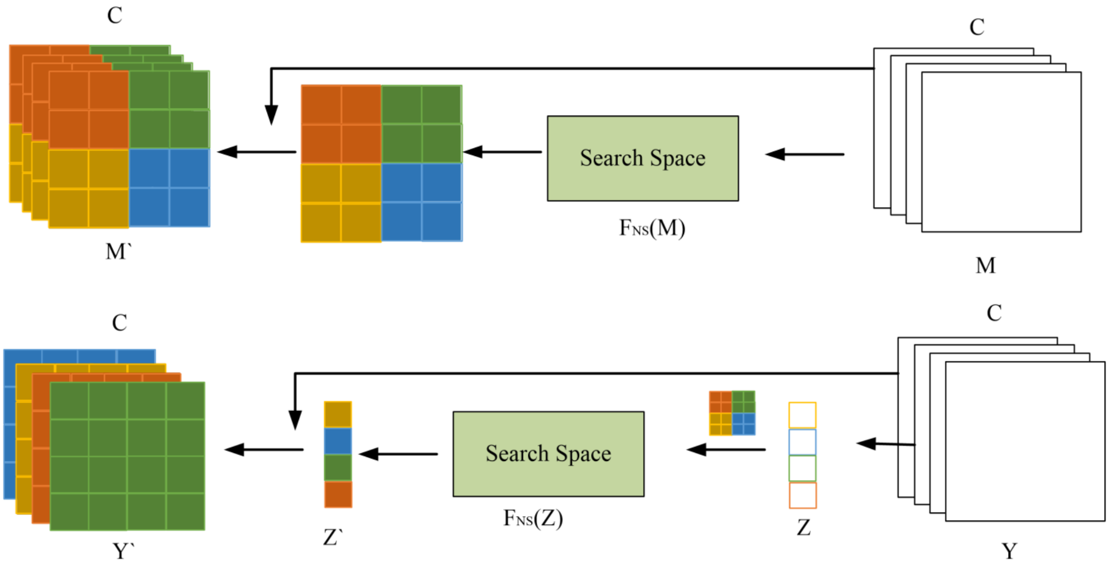


Figure 3.3: NAS attention module

3.3.2 NAS-Based Spatial and Channel Joint Attention Module

The design of deep neural networks, especially the combination of interdisciplinary networks, requires empirical knowledge and extensive experience. This is undoubtedly a time-consuming job. Therefore, we propose an automation model in this thesis, which stipulates the use of NAS to search and discover ships. NAS-SCAM consists of "room and attention channel" module, in which weighted output is achieved.

The input feature map is $M = [m_1, m_2, \ldots, m_c]$, which has width *W*, height *H*, and channel *C*, *M* is transformed into spatial weight map $n \in \mathbb{R}^{H \times W}$ by using one or multiple convolution operations and nonlinear operations $F_{NS(\cdot)}$ in NAS search space. In the search space, *n* is related to the spatial weighting information. Finally, we are use of the multiplication operation to fuse spatial weight map into the input feature map *M* and generate output feature map *M′* as shown in eq. (3.1).

$$M' = n \otimes M = \left[nm_1, nm_2, \ldots, nm_c\right] \tag{3.1}$$

Correct choice $F_{NAS}(\cdot)$ is the key operation that has great influence on the weight effect. However, there are many options, it is difficult to find the best one. Therefore, we recommend choosing $F_{NAS}(\cdot)$ by using the NAS to find suitable network structure.

In order to produce channel weighting information without changing spatial information, suppose the input eigenvalue map is $Y = [y_1, y_2, \ldots, y_c]$, $y_i \in \mathbb{R}^{H\times W}$, we take use of a global average pooling operation along spatial dimensions as shown in eq. (3.2) and generate a vector $z\in\mathbb{R}1\times1\times C$.

$$Z_i = Avgpool(y_i) = \frac{1}{H \times W}\sum_{p=1}^{H}\sum_{q=1}^{w} y_i(p,q) \tag{3.2}$$

NAS-SAM and NAS-CAM are employed for generating the feature maps, the two output feature maps retain important weights by using max pooling operations in the fusion. In order to better fit the attention search mechanism for the channels and spaces, a new search space has been created as shown in Table 3.1.

Table 3.1:The operations of NAS-SAM and NAS-CAM

	NAS-SAM	**NAS-CAM**
Zero (No connection)	√	√
Conv2D 1	√	√
Conv2D 3	√	√
Conv2D 5	√	√
Conv2D 9		√
Conv2D 15		√
Atrous Conv2D 3	√	√
Atrous Conv2D 5	√	√

Through the structure of NAS-SAM and NAS-CAM systems, we have selected multiple operations between the two nodes of the NAS search engine. Therefore, NAS-SAM and NAS-CAM operations are shown in Table 3.1. Regarding NAS-SAM, since we need information from the spatial dimension, we consider taking two-dimensional folds with the filters having different sizes to obtain information from the sensory field.

The NAS-SAM model is use of one-dimensional folding to extract channel information based on the average global pooling. In addition, NAS-SAM and NAS-CAM conduct a zero calculation, which indicates a lack of communications between the two nodes. The steps of the gradient-based algorithm are shown as follows:

Algorithm 3.1: Gradient-based algorithm

Input: Training

Output: Optimal network structure

- Step 1: Firstly, determine the number of model nodes.
- Step 2: While mixing operations, load all operations into the connection path nodes to form a neural multi-specialty network.
- Step 3: Apply different weights to each route so as to solve the discrete optimization problem and update the option weight combination of the hybrid operation at the same time.
- Step 4: Select the final network structure according to the possibility of hybrid operation.

In order to better take account of multifunctional combinations, we control the variance within a range where gradients can be computed continuously to optimize structure and operation.

Besides, in order to achieve the optimal attention module in the same network, we propose a synchronous search strategy to search each attention module separately. In general, the modules of the same structure are inserted at the end of each of the following units according to the attention module model. However, due to the convolution and pooling operations, the feature map has different semantic meanings at the different locations of the network. Therefore, while searching for different modules, attention will completely match the different points in the network, such as the upper and lower samples.

The synchronous search strategy supports the development and optimization of the only important modules for each of the proceeding and next blocks. Since each structure of the module is adjusted through the optimization of 120,572 continuous variables, the optimization in multiple directions produces a very suitable structure.

3.3.3 Loss Function and Evaluation Function

We employ a mix of losses in this thesis to better complete the assignment. As a starting point, the Binary Cross Entropy (BCE) loss function is employed to compute the loss between the prediction and the ground-truth, the cross-entropy loss function is applied to generate the class probability score, the Generalized Intersection over Union (GIoU) loss function is harnessed to forecast the bounding box. In order to maintain the rapid convergence and improve performance, the weights in the combined loss are set as c_{iou} = 0.05, g_{iou} = 1.00, and b_{ce} = 0.50, respectively.

The object detection probability, false detection probability, F1 score, precision, and recall are defined as eqs. (3.3-3.5) to assess the performance of the proposed deep neural network:

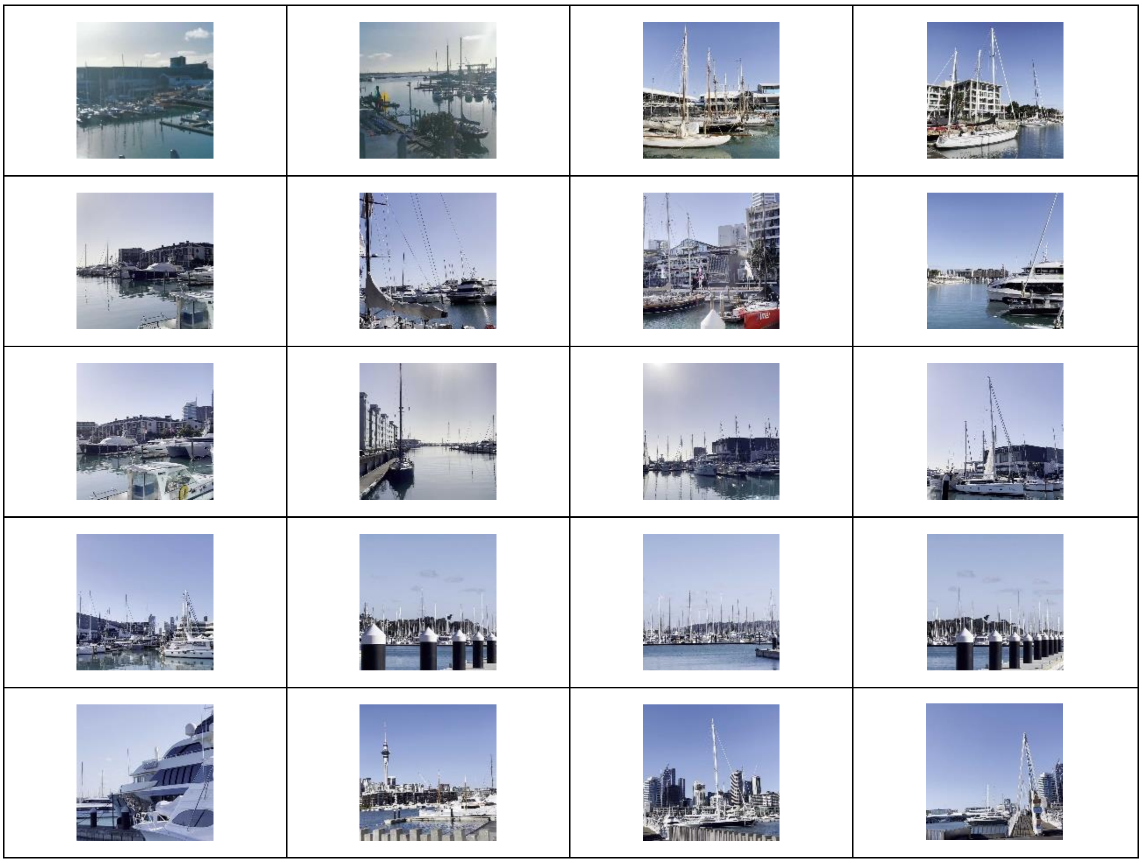
$$P_d = \frac{N_{td}}{N_{\text{ground\_truth}}} \tag{3.3}$$

$$P_f = \frac{N_{\text{fd}}}{N_{\text{total\_target}}} \tag{3.4}$$

$$F1 = 2 \times \frac{P_d \times (1 - P_f)}{P_d + (1 - P_f)} \tag{3.5}$$

where N_{td} is the number of true positive, $N_{\text{ground\_truth}}$ is the total number of ground truth, N_{fd} is the number of false alarms, and $N_{\text{total\_target}}$ is the number of detections in total.

Table 3.2: The image samples shot at local harbor



3.4 Kayak and Sailboat Detection Based on the Improved YOLO with Transformer

With the development of computer vision, its performance is getting better. The object detection is a classical task in computer vision. The most popular methods in object detection are YOLO and its family, which have been kept improving in the last 5 years. The versions of YOLO are already experienced from YOLOv1 to YOLOv5. However, in the field of ship detection, especially in the field of sailboat and kayak detection, there are few projects, or datasets. Therefore, in this thesis, we provide a dataset for sailboat and kayak detection. In addition, we evaluate the performance of YOLO models (Girshick et al., 2014) and Transformers (Vaswani et al., 2017) as well as the performance in the tasks of sailboat and kayak detection.

3.4.1 Unified Detection Model – YOLO

Joseph et al. launched YOLO in 2016, which is a modern unified object detection model based on CNN (He et al., 2017). A single-stage network was employed to carry out the layout of neural networks. The object detection problem is treated as a regression problem that directly deal with the class probability and location of the input images. Through the unified design, the object detection of YOLO is 10 times faster than that of other models. Regression models require a fixed-size input, however, when the input image is displayed, the network cannot accurately predict the number of visual objects.

In order to resolve this contradiction, YOLO has predicted a considerable number of objects and set a threshold to exclude low probability prediction. The principal structure and development of the YOLO model have been detailed in Chapter 2, so we will not reiterate it here. In this thesis, YOLO is very suitable as a unified model for visual object detection, we will improve and experiment on the basis of the YOLO model.

3.4.2 Backbone

The core difference between the YOLO model and other models like R-CNN (Girshick, 2015) is that YOLO only conducts one shot operation to get the probability of index. Regarding the regression problem, other models such as R-CNN decompose the original problem into “object detection” as a classification problem and “bounding box” as regression problem. Therefore, YOLO is more suitable for object detection than R-CNN. The main structure of YOLO is a group of convolution operations, and follows a fully connected layer (Redmon et al., 2016). On the other hand, the activate function of YOLOv5 model is leaky ReLU function, which contains negative values during the training process. It is more suitable for regression than ReLU. The loss function needs to consider the errors between prediction coordinate and the ground truth coordinate. Hence, a binary loss function will be useful in YOLO model, in this case, we select binary-cross entropy (BCE) loss function for YOLOv5.

- The main structure of Transformer is the encoder and decoder frame, which is an end-to-end learning algorithm to solve the sequence to sequence learning problem.

- The input is a sequence of data and the output is also a sequence of data. The input process is called encoder and the output process is called decoder, all memory will be employed to save content and context.
- Transformer has a multi-head attention method, which simulates human learning process. Attention method usually refers to computing convex combinations of content-based vector sequences, it tells that the weight itself is a function of the input.

The multi-head attention method can be considered as the integration of low-dimensional original attention layer which is always better than single head attention. Therefore, multi-head attention transformer will be implemented between the YOLO backbone and the fully connected layer.

The main idea of Transformer is attention, whose basic frame is the encoder-decoder structure. Therefore, to combine Transformer with YOLO model, we need to split the original YOLO model into two parts: Convolution backbone and the fully connected layer. The convolution backbone of YOLO is for extracting the image features, the fully connected layer is for generating output. In order to combine Transformer and YOLO models, we need to connect convolution backbone with the encoder, and connect the decoder with the output of fully connected layer.

YOLO models were employed for visual object detection which only needs to "look once" and inferences the index with probability directly. However, before YOLO was proposed, region-based fully convolutional networks (R-FCN) had been widely used to inference visual object first and then predict the suitable index location. A moving 2D window is employed to search on the feature map to find the most suitable position of visual object. R-FCN is a typical method of visual object detection, with ResNet-101 being its convolution backbone structure.

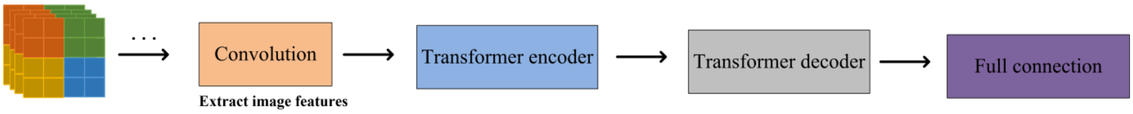


Figure 3.4: The backbone of YOLO with Transformer

3.4.3 Training Method

As deep learning methods become stronger and stronger, the max depth of neural network and the depth of backbone become stronger and deeper (Dahl et al., 2012). Therefore, the optimizer (e.g., Adam) needs a few training samples to avoid overfitting and obtain the theoretical target solution (weights of the networks). ImageNet is a popular and useful dataset in computer vision, which contains a plenty of images (Maurya et al.,2021).

After BERT model came out, pre-training has become much usual (Devlin et al., 2018). The main idea of pre-training is to replace the random weights of feature extraction layer with a trained weight. This method can be employed in the similar task. In general, pre-trained weights can reduce the total training time for a specific task.

This method is implemented after getting a completed group of weights of feature extraction layer. By using the dataset of specific tasks, training the fully connected layer and making a little bit change in feature extraction layer, this method is called fine-tuning (Zhang & Hu, 2021). Based on this method, we took use of it in sailboat and kayak detection to reduce our model training time.

3.4.4 Uniform Blending

As we have seen, if averaging two models together, the generalization error will always be equal or less than the weight sum of each single model. Therefore, the blending method will be chosen in the end of this project to improve the performance of model. The pseudo code of this model is listed as Algorithm (3.1) and Algorithm (3.2).

Algorithm 3.1: Training a model

```
Input: Training set; $N$: Number of total training images; $CFG$: Initial parameters; : $Fold$ number; lr: Learning rate
Output: Optimal Model: $M^*$; Out-of-fold Prediction set: $P_{\mathrm{oof}}$
Initial random state and model weight $W_0$;
repeat
    Set random seed $R$;
    Divide dataset into training set and validation set,
    Load the YOLO model structure by using CFG;
    Load the pre-training weight.
    repeat
        Set Adam optimizer and a stable learning rate $lr$;
        Train the model;
        Compute the target loss $Loss_{\mathrm{local}}$ by using $BCELoss$
        Update $Loss_{\mathrm{best}}$ if $Loss_{\mathrm{local}} < Loss_{\mathrm{best}}$
        Update saving if $Loss_{\mathrm{local}} < Loss_{\mathrm{best}}$
        Loading valid set and compute valid loss;
        Save valid loss with best model as out-of-folder result set Poof;
    until Epoch times OR $\mathrm{Loss}_{\mathrm{best}}$ has no change for 3 epochs.
until $fn$ times
```

Algorithm 3.2: Transformer modeling

```
Input: Input image;
Output: Out: Model Out;
Image resize from $D_{\mathrm{tr}}$ to;
Take Batch Normalization from to $BN_{\mathrm{D}}$;
Implement $BN_{\mathrm{D}}$ Linear Transform and get $L_1$;
Reshape $L_1$ to multi-head from and select 3 dimension: $q, k, v$;
$q \times k$ and implement transpose, gives $a_1$;
Take softmax to $a_1$ and gives $a_2$;
Implement Dropout at $a_2$ and gives $a_3$;
$a_3 \times v$ and transpose back to original image shape $a_4$;
Take the linear transform to $a_4$ and get $a_5$;
Dropout $a_5$ and gives $a_6$ as attention output;
```

```
Implement Drop path for $a_6$ and gives attention layer out: $a_7$;
Add original input and $a_7$ gives output: $Out_1$;
Take Batch Normalization to $Out_1$ to get $m_1$;
Send $m_1$ into full connected layer 1 and get $m_2$;
Use GELU as activate function to $m_2$ and get $m_3$ ;
Dropout $m_3$ and gives $m_4$ as MLP layer 1 output;
Send $m_4$ into full connected layer 2 and get $m_5$;
Dropout $m_5$ and gives $m_6$ as MLP layer 2 output;
Implement Drop path for $m_6$ and gives MLP layer out: $Out_2$.
Combine two layers together, then gives $Out = Out_1 + Out_2$.
```

3.4.5 Implementations

In this section, the model is implemented with details. It will involve the training process and the method about how to get the best parameters of each model. A pre-training method will be employed for this task. The ImageNet-1k pre-training weight is implemented as the initial weight of feature extraction network in YOLO models, which will involve convolution operation to easily extract the feature maps of visual objects. We start the first run with epoch 10, learning rate 0.0001 and the binary cross-entropy loss.

As all data samples are tagged, all training set needs to review again to make sure there is no wrong labels. The data samples are captured by using mobile cameras, whose original resolution is 4K×4K. It will take up a huge amount of computer space and computing power, thus, we set the input size of YOLOv5 model as 640×640, which will take the bytes of size in GPU (Park et al., 2021). Correspondingly, the input samples are also resized as 512×512 and 256×256.

The modeling environment is with a 6G RAM GPU. From the testing result, we determined the input size, which shows 512×512 as the input size will generate a better performance.

Figure 3.5 shows 8.00×10^{-4} is the best learning rate for YOLOv5 model in this group of experiments. Other training methods are set as this group of parameters. Therefore, we additionally test the datasets with sizes 640×640 and 256×256. The experimental results show that the learning rate is optimal under various sizes of input images.

Table 3.3: Training progress with various input sizes

Datasets	Quantity	Image sizes	Epochs	Time (h)
S&K-1000-Original	2,787	640×640	3	5.7
S&K-1000-Cleaned	2,749	512×512	3	4.5
S&K-1000-Cleaned	2,749	256×256	3	3.6

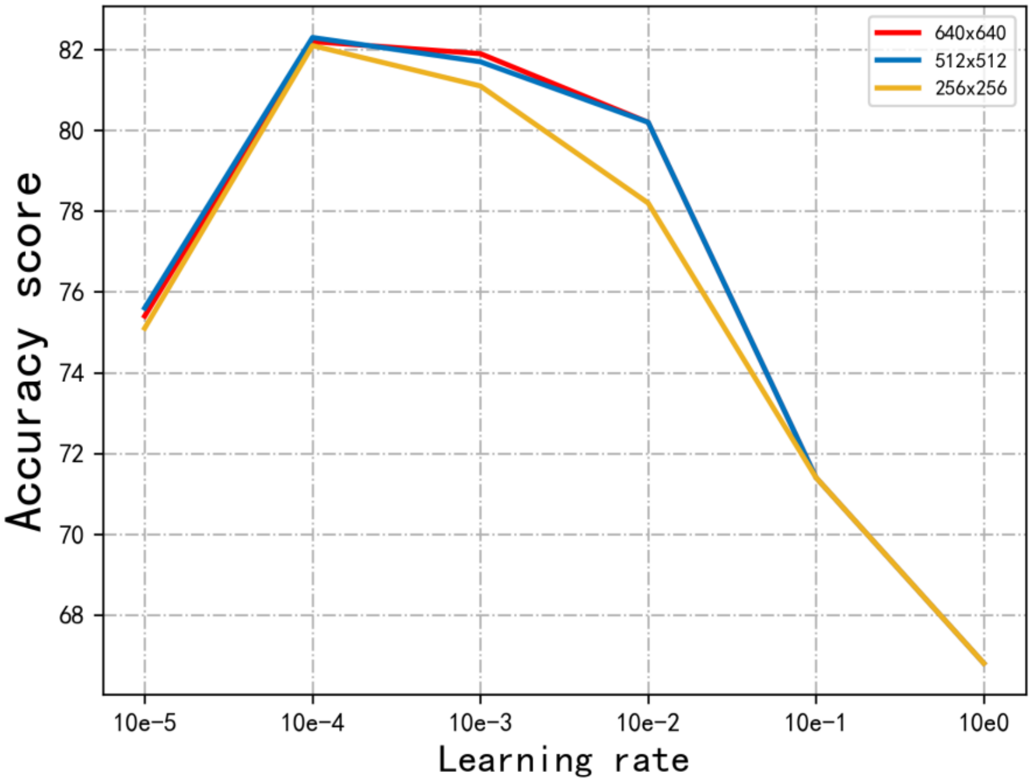


Figure 3.5: The results of learning rate search by previous input size

To verify the proposed learning rate, the number of epochs is increased up to 50. For the resource saving purpose, this process is implemented by using specific rates such as 0.8, 0.1, 0.08, ..., 0.00008, 0.00001.

We also test those popular loss functions to replace the cross-entropy function. The log loss, exponential loss, hinge loss and cross-entropy loss functions all are applied to calculate the accuracy, as this problem is much like a binary classification. According to the training results, the cross-entropy function is the best one in this problem with YOLOv5 model.

Table 3.4: Cross-validation scores with different loss functions

Loss functions	**Epochs**	**Minimal Losses**	**CV Scores**
Cross Entropy	5	0.0018	0.2314
Log loss	5	0.0033	0.3124
Exponential loss	5	0.0037	0.2928
Hinge loss	5	0.0089	0.2882
Categorical Cross Entropy	5	0.0012	0.2135

In Table 3.4, the training process was performed as convergence if the number of epochs is set as 5. If we set all methods with the epoch 5, it will save computing time. Thus, we use the same parameters in the Transformer-related model.

3.4.6 Ensemble Learning

YOLOv5 is determined by its baseline, with input size 512×512, learning rate 0.0008, and the binary-cross entropy function for the loss computations. To further improve the performance of the baseline model, we ensemble the models to test how to get a better score. The ensemble method has the voting and blending operations, where the blending method is for calculating the probability(confidence) of each visual object, the voting is to decide which object is output.

Each model will eventually output a group of prediction results with probabilities. To blend the results of object detection, we need to count the overall indices, and set probability 0 to each index with other objects. It is essential to make sure that all indices have a series of predictions with “Kayak”, “Sailboat”, and “Other Boats”. Then, ensemble learning will calculate the average of each model and give every model a weight. The next step is to take the weighted mean as the final probability. In order to find the best weight of each model, we search for the best value by using the cross-validation prediction result.

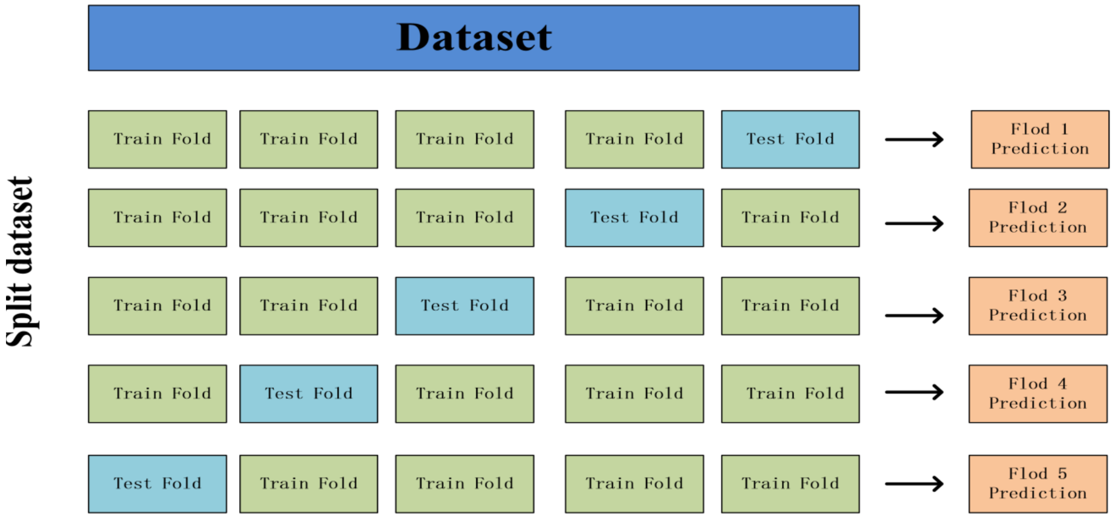


Figure 3.6: A valid predicting label which was used for calculating the losses with true label

Table 3.5: Ensembled results

Model Structure	**Number**	**CV Score**	**Test set Score**
YOLOv5	1	0.2298	0.3014
Transformer	2	0.2043	0.2833
1+2 Ensemble	3	0.1989	0.2798

After obtained the probability of each model, the next step is to vote these models. Each model has the votes as same as its probability, the class will be determined through the votes. The final output will be marked as its label. To increase confidence, if the probability is too low, it will be regarded as a wrong prediction and removed. Table 3.5 shows the ensembled results which reduce the cross-validation errors and attain a better outcome.

Chapter 4

Results

The main content of this chapter is to describe the experimental setup and demonstrate the experimental results. In this chapter, we will also brief the limitations of this project.

4.1 Experimental Setup

The main task of this chapter is to introduce an object detector, test the output results of the ship detectors based on the datasets. These tasks and the conditions are related to the criteria for selecting experimental methods.

4.1.1 Faster R-CNN

The three ship detectors are R-CNN, YOLOv2, and SSD with Transformers. Fast R-CNN is the most advanced detector. YOLOv2 makes use of the deep net to achieve the real-time testing goal of SSD by using a hierarchical structure characterized by a pyramid net while ensuring accuracy and computational speed.

Faster R-CNN model was trained by using nine region proposals for each sliding window in the RPN. These nine region proposals include the scales 64, 128, 256 as well as the aspect ratios 1:1, 1:2, 2:1. The aspect ratios are adopted from those existing methods (Ren et al. 2015). The SuperView dataset is with larger scale than the multiobject class datasets.

Fast R-CNN resolves the counting problem based on R-CNN and SPP nets. In order to improve speed and accuracy, it takes advantage of pooling operations and a small number of samples. However, it still needs a long time due to using the selective search method. In contrast, it is faster and more creative to use CNN and extract the region proposals through region proposal network (RPN).

Fast R-CNN detects visual objects according to the proposals obtained from the RPN network. This increases the network speed to detect visual objects. The initial RPN net usually produces nine anchors in one place, with 3 learning rates, the aspect ratio between width and height is less than 1.0. Based on our experiments, a confidence score threshold 0.6 has proven to be appropriate for this problem which was taken into account during experimentation.

In order to improve generalization, dropout is included in the classifier for visual object detection. Similar to the activation function in Faster R-CNN architecture, ReLU function has been employed. The regression layer takes use of a simple linear activation function, so all ships are marked as bounding boxes (i.e., BBox).

4.1.2 YOLO

YOLO and YOLOv2 algorithms were employed to visual object classification and positioning through deep neural networks. This makes ship detection becomes a regression process at the final stage of visual object detection. YOLO algorithm is much faster than region-based CNN. However, its detection rate is relatively low, thus with a lot of object detection errors.

Based on a dataset with a small number of samples, this needs data augmentation. If an image is chosen, it needs three operations: Horizontal flipping, vertical flipping, and Gaussian noising. YOLOv2 model was trained with Adam optimizer and a learning rate. There are no pre-trained weights in visual object detection. The initialization is conjunction with ReLU function or Leaky ReLU function. The confidence probability is subject to the sigmoid activation function $\sigma(x)$.

YOLOv5 and YOLOv4 are relatively new, YOLOv5 is only one month after YOLOv4. Therefore, the structures of YOLOv5 and YOLOv4 are very similar. YOLOv5 has not changed too much compared with the previous YOLOv4. It is very tough to find relevant publications related to YOLOv5.

Using 3D convolutional neural networks, YOLOv5 was developed to leverage temporal information. The feature component of YOLOv5 is unchanged. The last layers are utilized to extract spatiotemporal feature of visual objects with small, medium, or big size. Figure 4.1 depicts the detailed structure, each block is detailed as follows.

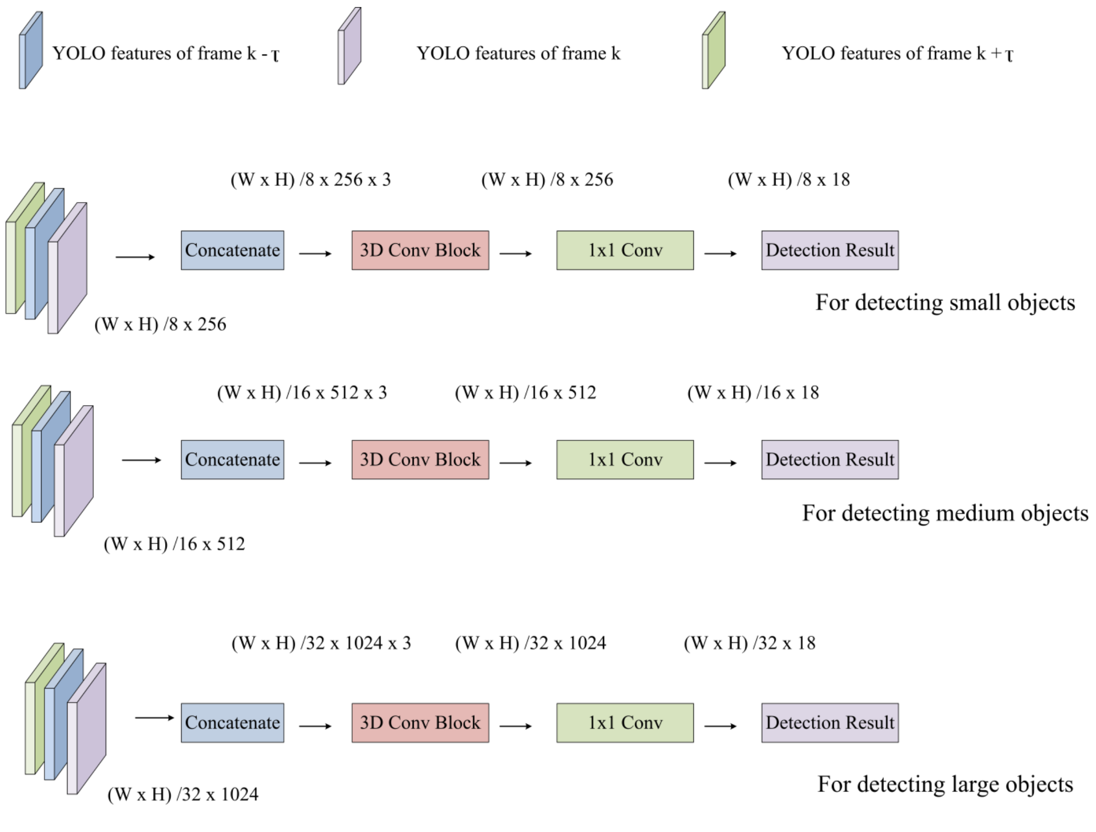


Figure 4.1: The proposed YOLOv5-temporal structure

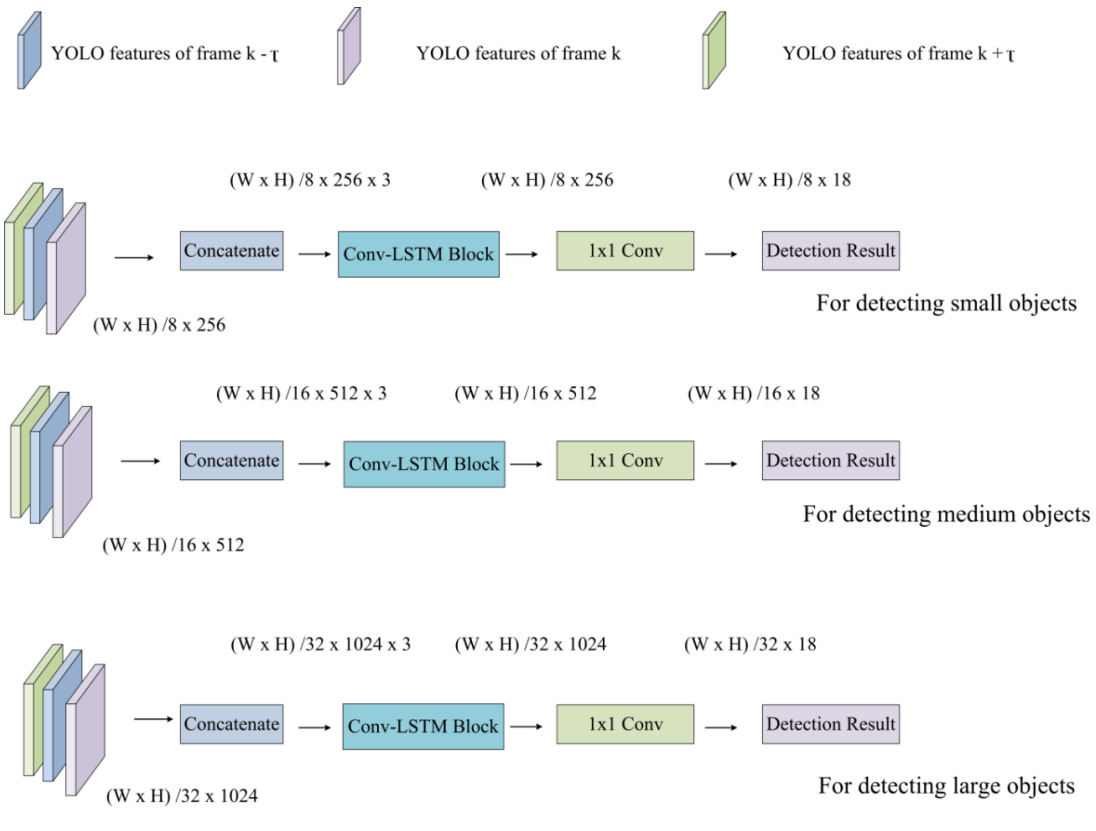


Figure 4.2: The proposed YOLOv5-LSTM structure

By using Convolutional LSTM, YOLOv5-LSTM model was developed to leverage temporal information. The feature extraction of the YOLO stays unchanged. In YOLOv5, the last layers are adjusted during regression to extract spatiotemporal properties of visual objects for small, medium, and big size. Figure 4.2 depicts the details of this structure.

YOLOv5 was developed based on Python, rather than C as previous versions. This simplifies the useability of integration and installation. Because they were developed with two different languages, it is a difficult task to compare YOLOv4 and YOLOv5 from the viewpoint of crossing platform development.

4.2 Results of Sailboat Detection Based on Deep Learning Models

4.2.1 Experimental Setup

(1) Datasets

Typically, in this thsis, visual objects in an image are marked as boats or backgrounds. Although different types of ships have the same elements, such as deck and stern, most of ships still differ obviously in shapes and appearances. The differences make it difficult to find specific class of ships.

Several datasets were utilized to train deep nets throughout this thesis. The offer of this thesis is to either pretrain or train the deep neural network models. Ship-specific datasets are utilized to fine-tuning a network.

The first dataset presented in literature was to match a ship silhouette as database records. It is made up of a variety of ship images. Each image is labeled with the ship classes and a per-pixel segmentation. For each image, segmentations are supplied that have been annotated by three individual human operators. Consequently, each image has three slightly distinct segmentations. A majority vote was employed to create a single ground truth of segmentation map for each image.

The collection is made up of ship images captured at various view angles and distances, with various backgrounds. Because the ship backdrop offers minimal information about the ship classes and brings substantial noises to the image, we removed this randomness by cropping a region of the image that only includes the ships. The goal of this thesis is to crop the ship images without removing any ship-related pixels. Because the ship has well-defined borders, we firstly conduct edge detection based on the first image.

Because the ship borders are precisely defined, the aim is to identify the rows and columns of pixels in the image corresponding to these edges. As a result, we add the pixels along a row in the image and compare them to the sum in the next row. If the values in the two rows differ significantly, this indicates an edge in that row. Edge detection identifies random artifacts because of random backdrop. We find the sum across a few rows, which overcome noises and make the approach resilient against mistakes. We crop the images after locating the row and columns of pixels where there is an obvious change in edge detection.

In this thesis, we firstly created a real dataset using the videos from surveillance cameras. It was used to evaluate the effectiveness of our proposed deep learning models. Secondly, we converted the video frames from America's Cup in the past three years to a standard size. In the training dataset, 1,484 images with labels were included for model training, 348 ships were contained for the test. For the training dataset, we select the image resolution 512x512. Ship patches were sent to the deep network as the training input. We tackled the test images in the same way to ensure the fast and correct operations of deep learning models. Our dataset also includes the frames from the America's Cup videos, as shown in Table 4.1.

(2) Dataset splitting

In order to evaluate the performance of a network, we must firstly create a test dataset. This guarantees that the performance is unaffected by samples for network training. While most datasets include a wide range of images, a huge dataset comprises several subsets of images, the images of the same vessel are also included. Each image is labeled with a unique ID that corresponds to the ship in the image. Splitting the dataset based on this ID ensures that the same ship appears in both the training set and test set with slight difference. A difficulty for this splitting is that numerous ships might belong to the same class, that means, they have the same classification output. To correct this, a division based on classes was created.

(3) Experiment setup

We carried out our experiments by using PyToch. The baseline of the proposed net is YOLOv5. Regarding NAS search, the process is appended at the end of each downward or upward sampling block. The total iterations are 120, where W is updated as 120. After the optimization training, we restored the network according to the training level. In the recovery process, the total number of iterations is 300, the learning rate is 1.00×10^{-3}. After the verification, we retained the most powerful model and treated it as the best one for our experiments.

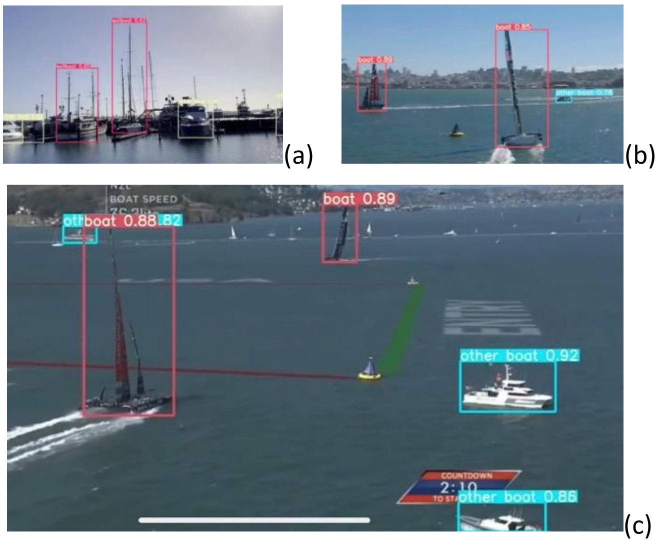


Figure 4.3: The screenshots of demo videos (a), video (b) & video (c)

In Figure 4.4, we show various results of sailboat detection in our test. Figure 4.4 (a) is a video from our harbor, Figure 4.4 (b) is a video from America's Cup 2021, Figure 4.4 (c) is a video to detect various sailboats.

4.2.2 Experimental Results

The small ship dataset is gathered in two ways. Firstly, we captured the images of real-world tiny ships (positive samples) and images without ships (negative samples) near the wharves. At the same time, because the images of small ships acquired is not enough for us to conduct training and testing, we augment the positive training 99 samples using tiny ship images gathered online. The samples were designated as the baseline dataset. Table

4.1 shows the examples of obtained samples from the baseline dataset.

In order to complete this task, the suggested dataset GMWGAN-GP was utilized in the tests to create positive samples (i.e., fake images of small ships). LabelImg, an open-source program, was applied to annotate the actual data as well as the samples created by GMWGAN-GP. The image name, object classes, coordinates and sizes of the bounding boxes or rectangles are all included in the created annotation labels.

Theoretically, the number of images generated through the network is unlimited. However, if the number of images is too large, the quality of image samples cannot be guaranteed, which will eventually affect the detection ability of the ship. Therefore, in the experiment, the basic dataset was collected, the positive samples are split into several groups to study the relationship between the proportion of positive samples and the test results.

As part of this work, we have made horizontal and vertical comparisons to better compare the effectiveness of our proposed model with others. Horizontal matching means that we compare the properties of the shared. On the other hand, we compared the characteristics of different models based on the same dataset, which represents a vertical comparison. Firstly, we look for models in a centralized search of public data. In order to verify the effectiveness of our proposed NAS-SCAM and synchronous search strategy, we compare the characteristics of the proposed model.

Table 4.1: Comparisons of ship detection based on a public dataset.

Models	$\mathrm{P_d}$(%)	P_f(%)	$\boldsymbol{F1}$
Baselines	71	18.5	0.75
NAS-SAM + Baseline	72.49	19.30	0.77
NAS-SAM + Baseline	73.50	20.50	0.78
NAS-SCAM + Baseline	72.50	18.60	0.76

Table 4.2: Comparisons of ship detection based on our own dataset

Model	$\mathbf{P_d}$(%)	$\boldsymbol{P_f}$(%)	$\boldsymbol{F1}$
Baseline	67.00	17.00	0.69
NAS-SAM + Baseline	68.49	18.30	0.72
NAS-SAM + Baseline	66.00	16.50	0.70
NAS-SCAM + Baseline	71.00	19.60	0.77

In Table 4.1, we see that NAS works more efficiently than basic types of machine learning algorithms without attentions. In addition, we see that the channel-based monitoring mechanism is better than the local monitoring mechanism. In order to further verify the characteristics of our model from the perspective of generalization, the accounting-based model is verified based on actual image acquisition.

In Table 4.2, the model based on structure search is more perfect than the typical CNN mode. In the comparative experiment, because other parameters remain unchanged, we see that the effectiveness of the model is reduced by comparisons, thus we eliminate the searching mechanism of neural structure. This confirms the effectiveness of the proposed method.

4.3 Results of Sailboat Detection Based on the Improved YOLO with Transformer

4.3.1 Experimental Setup

The structure of YOLOv5 model is identical to that of the previous YOLO series, which is split into four parts: Backbone, input, prediction, and neck. The primary structure of YOLOv5s is seen in Figure 4.4.

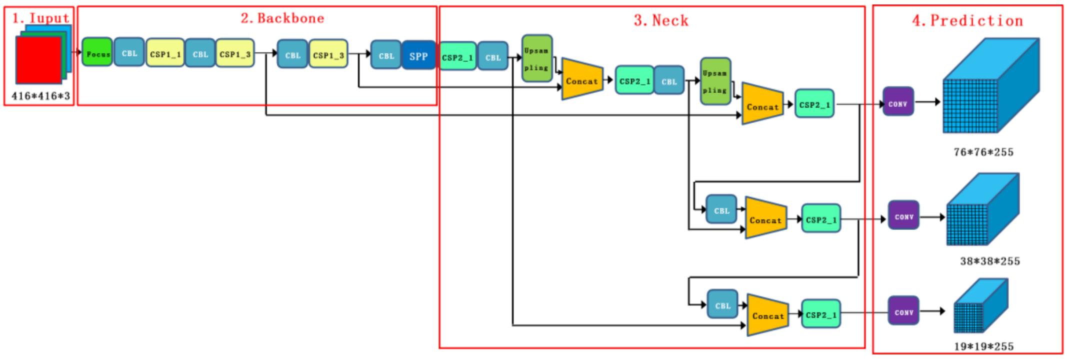


Figure 4.4 The main structure of the YOLOv5s model

The input component is capable of performing data augmentation, adaptive anchor frame computations, and adaptive image scaling. The focus structure, which can complete convolution and slicing operations, the CSP structure, which improves the feature network's learning capacity, is used mostly in the feature extraction section.

Because the number of convolution kernels in the Focus and CBL of various networks varies, the model performance can vary by adjusting the network breadth and depth. The neck section employs PAN and FPN structures, applying information retrieved from the backbone part to improve network feature fusion. The output layer is separated into convolutional layer that is generated by using the loss function which is subject to the maximum value suppression process to get the prediction outcome.

The datasets of this thesis are separated into two groups: Self-created datasets and public datasets. The public dataset is the SeaShips dataset (Shao et al, 2018), which contains images from surveillance camera installed along coastline, as well as images from every frame of the surveillance videos. The dataset was compiled from ships on water. The mosaic enhancement approach was applied to randomly choose four images for random scaling as well as randomly distribution for splicing, considerably enriching the detection dataset, particularly due to the random scaling introduced many small objects, making the network much resilient.

In Figure 4.5 (a), we depict a map of anchor and the intuitive situation of data labels, as well as an overall analysis of the object size and position on the labelled image as shown in Figure 4.5 (b), a relative size map of visual objects as shown in Figure 4.5 (c).

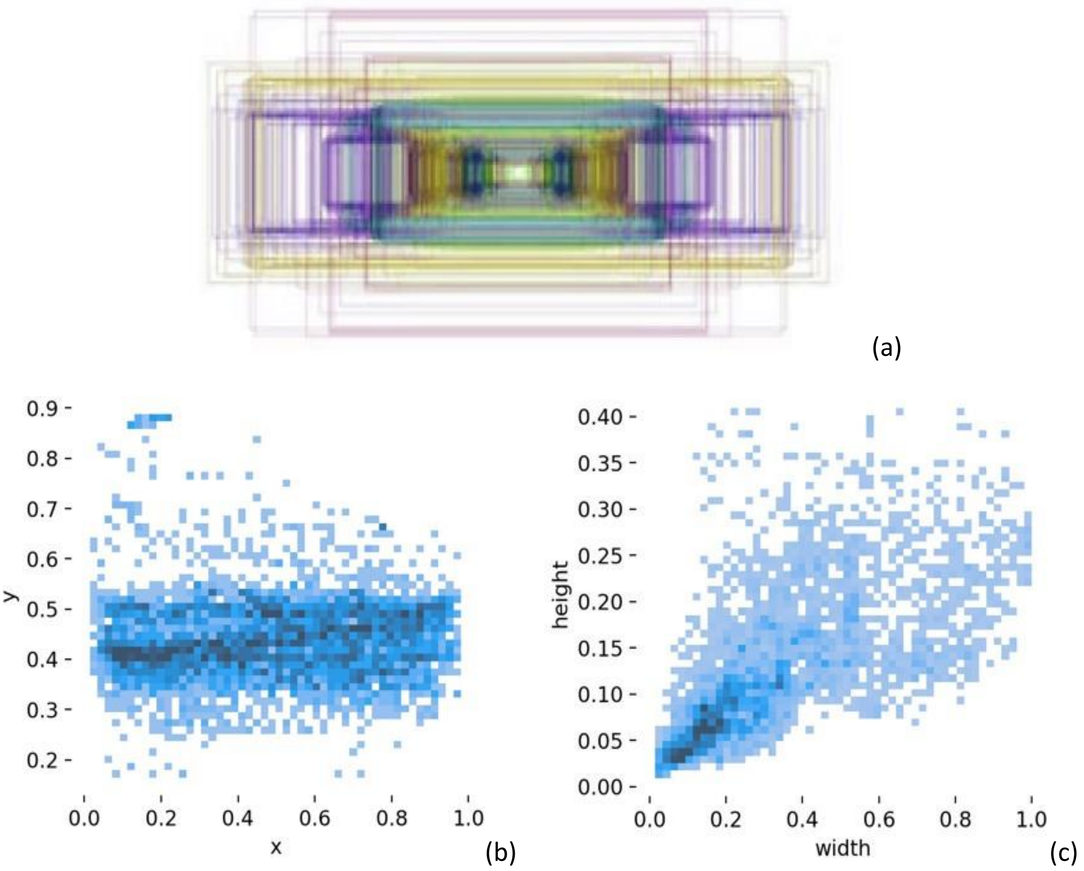


Figure 4.5: Statistical results of visual samples, (a) anchor distributions, (b) normalized object locations; (c) normalized object size in width and height

Figure 4.6 (b) shows that the coordinate origin was placed at the bottom left corner, the relative coordinates consist of ordinates *y* and *x* which were applied to describe the relative location of visual objects. Figure 4.6 (c) indicates that the object width generally filled 25% of the image width as well as the object height mostly occupied 58% of the height of image.

There was a significant gap between the distribution of the dataset and the initial set of regional candidates, the dataset contains a wide range of visual objects with varying sizes, which leads to unbalance of object distributions. As a result, the initial objects were clustered, the receptive field and loss functions were enhanced.

4.3.2 Experimental Results

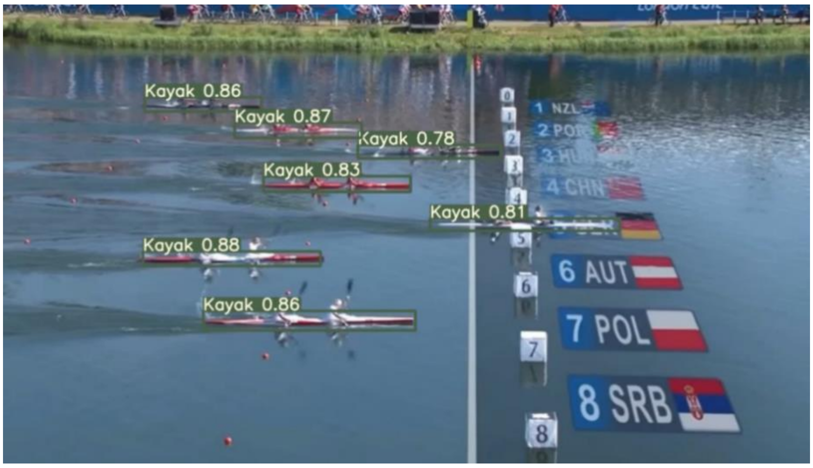


Figure 4.6: The result with a high confidence and the output with a high probability

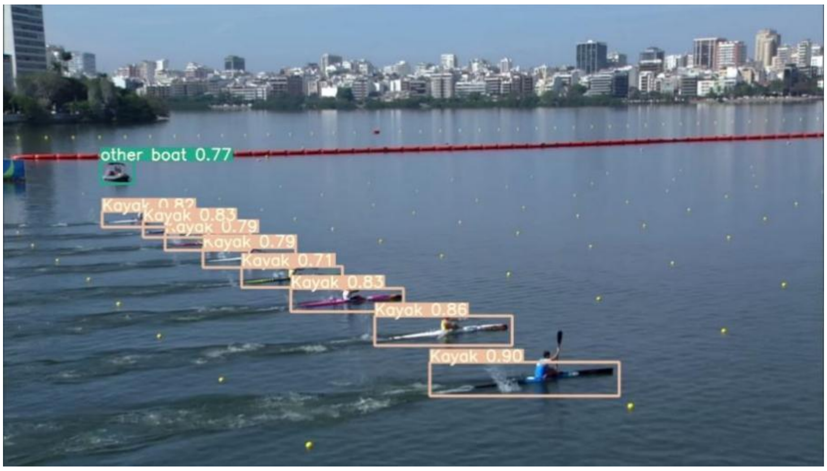


Figure 4.7: The test image from the Olympic Games

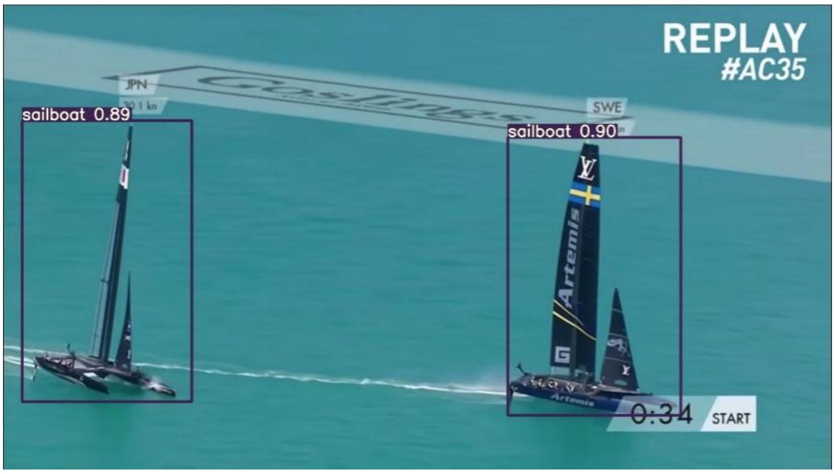


Figure 4.8: The output of sailboat class of the testing result

Figure 4.6, Figure 4.7, Figure 4.8 show that the index is correct, but it gives a probability around 80%. In this case, the final accuracy will be good, but the cross-validation error and generalization error will be influenced by this “un-confidence” probability.

The accuracy of the proposed method is utilized as the assessment metric. Based on all datasets, we additionally report the Intersection over Union (IOU). IOU is defined as the ratio of the area of overlap to the area of union which represents the intersection of the truth ground bounding box and the bounding box of the regression result.

In addition, for a more thorough comparison, we present the recall rate across all datasets. It indicates that the object detection results of the proposed method based on supplemented datasets are all greater than those based on the basic dataset. On enhanced dataset, the detection results of the proposed method achieved the highest accuracy (97.2%), IOU (84.2%), and recall (92.3%). As a result, in the next part, the enlarged dataset is utilized as the training and test dataset for comparisons. Table 4.4 shows that the results of different cases based on the same proposed approach.

Table 4.3: Comparisons between different results

CASE numbers	CASE accuracy
1	1.000
2	0.972
3	1.000
4	0.999
5	0.998
6	1.000
7	0.999
8	1.000

In Table 4.3, we see all the research cases show high accuracy based on same method. Furthermore, each parameter of YOLO5x model fluctuated significantly in the 0~50 rounds, which indicates that the model was highly unstable for detecting tiny objects.

Figure 4.9 depicts the precise circumstance. In the two images, the abscissa represents the epoch, while the ordinates represent the loss and mAP values.

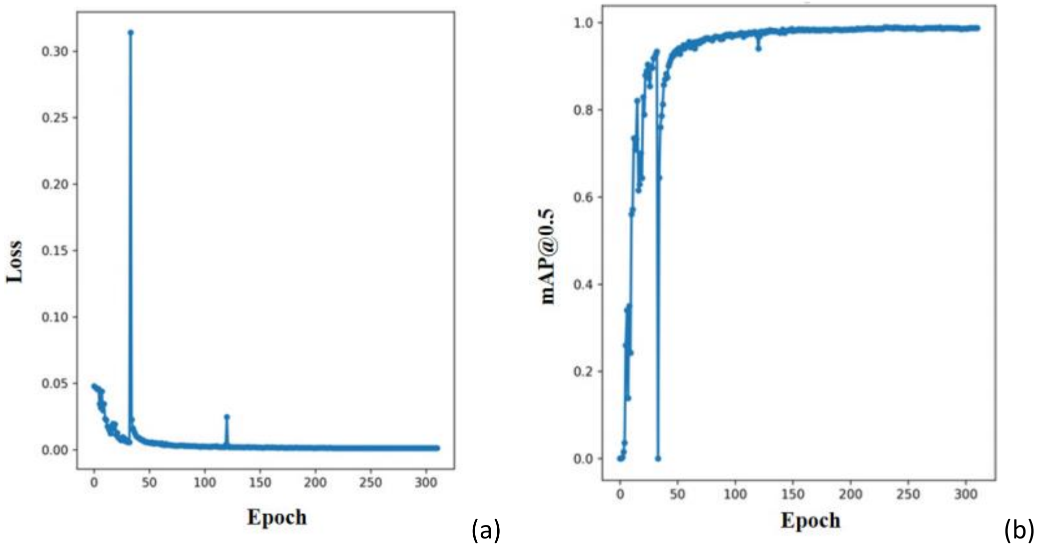


Figure 4.9: YOLO v5x training results: (a) Loss function curve of YOLO v5x, (b) YOLO v5x mAP@0.5 curve.

The durations of visual object detection using YOLOv5l and YOLOv5m models were too long which could not meet real-time needs; YOLOv5s model had a quick detection speed and met real-time requirement. One explanation is that its poor accuracy might be the reason why the model is ineffective at recognizing tiny objects, and the output has biase.

4.3.3 Analysis and Comparison of the Improved Model

Figure 4.10 shows the PR curve of YOLOv5s model. The modified model obtained strong outcomes for detecting all kinds of ships, the AP rate for all ships reaches 99.6%. The confusion matrix is depicted in Figure 4.11.

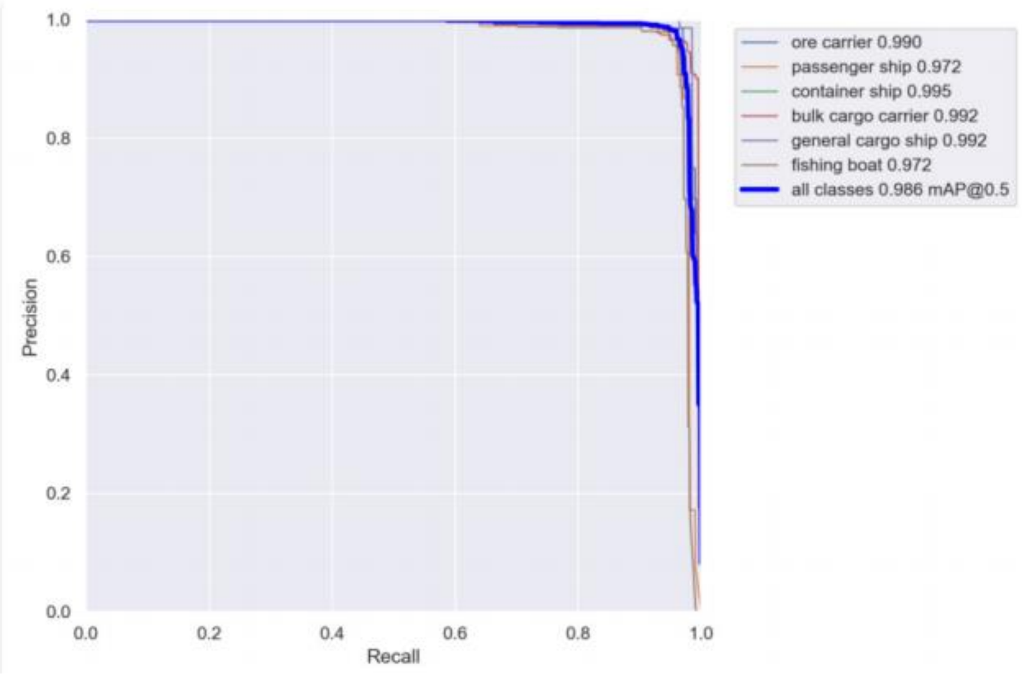


Figure 4.10: PR curve of YOLOv5s

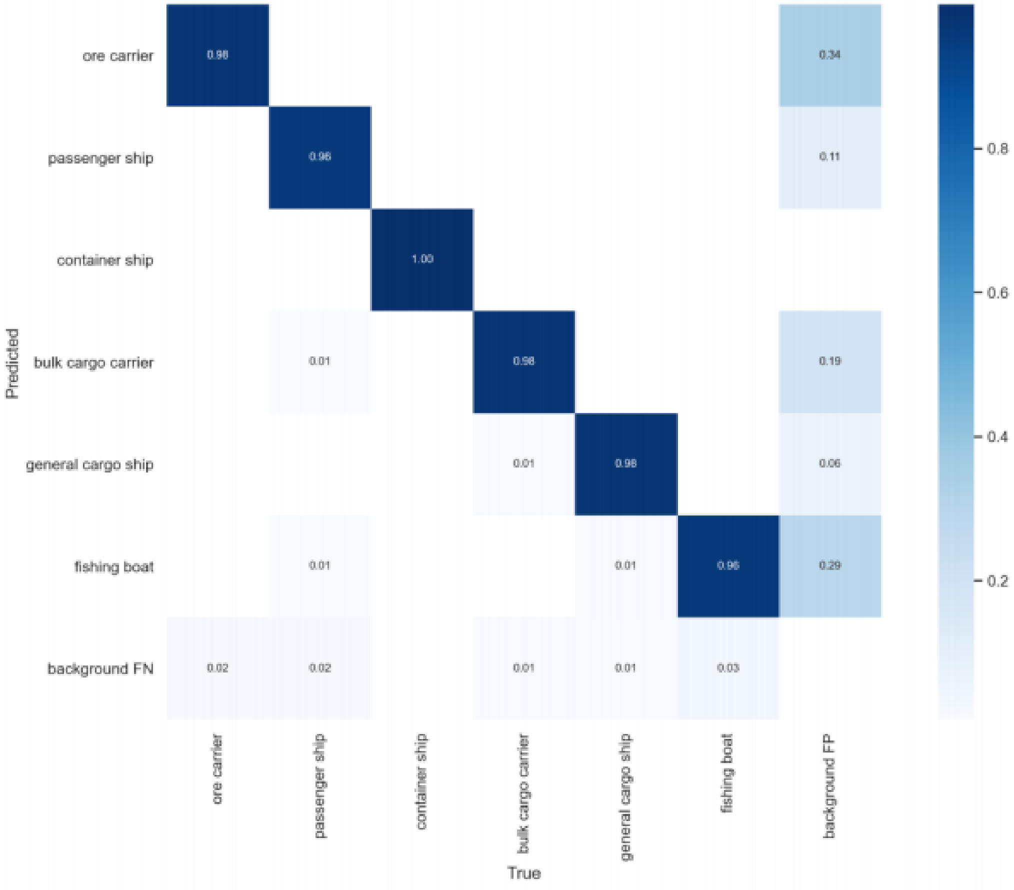


Figure 4.1: Confusion matrix of YOLOv5s

In next step, we increase the capacity for tiny object detection according to the comparison. Table 4.4 shows the increase of accuracy after we include this tiny object detection method.

Table 4.4: Accuracy increase after including tiny object detection

CASE number	**CASE accuracy**
1	0.831
2	0.543
3	0.279
4	0.777
5	0.862
6	0.242
7	0.344
8	0.92

As the results shown in Table 4.4, the capacity of the proposed algorithm to recognize tiny visual objects was much enhanced, the error rate was lowered. Although the time of ship detection rose 2.2 ms, the mAP improved up to 4.4% than the original method, which indicates that the enhanced network could fulfill the demands of accuracy and outperforms YOLOv3 and YOLOv2.

Chapter 5

Analysis and Discussions

In this chapter, experimental results are evaluated and analyzed. Comparisons of the results under various conditions will be conducted.

5.1 Analysis

In this thesis, we proposed YOLO models for sailboat detection, we introduced the monitoring mechanism of automatic search: NAS-SCAM-YOLO. This algorithm needs less time on improving the accuracy of ship detection. The sailboat images can distinguish them from other vessels. Our basic idea is to select our own extracted vectors and optimize the attention of the model connection while maintaining the rapid prediction of the regression algorithm, so that the whole network can better filter out its own vectors for subsequent verification. At the same time, in the feature extraction networks, the observation mechanism improved integration method of features described in NAS-SCAM-YOLO.

The next step of this project is to investigate in-depth visualization of the attention mechanism and present an intuitive representation of the visual features within the model.

In the project, we implement kayak and sailboat detection, and successfully mix YOLOv5 and Transformer Backbone together for visual object detection. For each model, we conduct experiments to find the best parameters such as input size, learning rate, and the best loss function. Finally, we ensemble these models and get a model with a less cross-validation error and generalization error. The next stage of this work will keep looking for other suitable backbone structure to expect a better result.

As a result, in this thesis, we discuss the best result of YOLO models and successfully improve the YOLO performance by using ensemble method. In the next stage, in order to improve the model performance, we are able to:

- Implement more models and ensemble them together.
- The dataset has repetitive images, so cleaning dataset could also uplift the performance.
- In blurry image, the confidence will reduce. Hence, we will conduct data augmentation, e.g., blurring, sharpening to increase the robust of the training models.
- After the data augmentation, TTA can also be considered in the test set.
- According to the first stage process, the general structure of object detection is a

convolution backbone for image feature extraction, the output is a feedforward neural network. In this case, there may be other methods that could be used to replace the attention model.

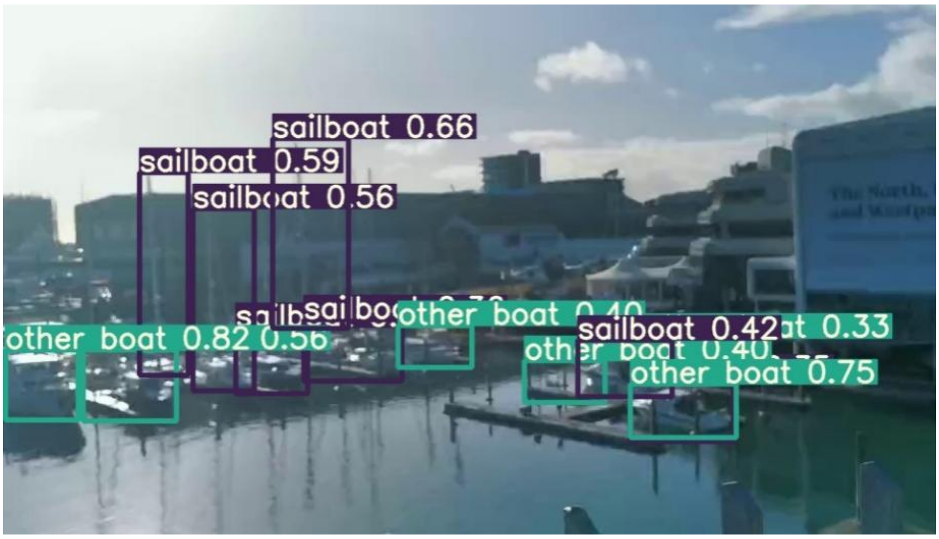


Figure 5.1: The condition with blurry image with all types of boats

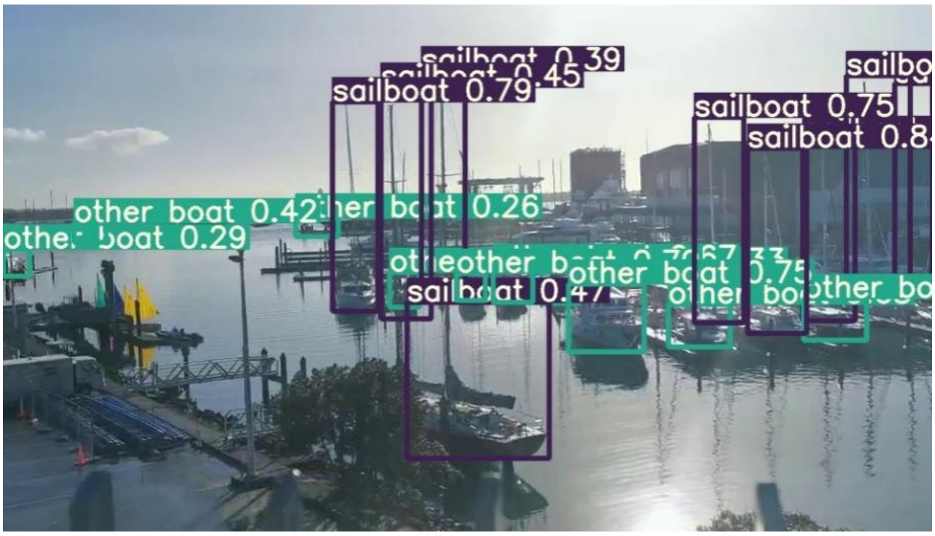


Figure 5.2: A scene with overlapping goals

5.2 Discussions

5.2.1 Comparison between Models

Model selection is frequently influenced by the speed of object detection. Depending on picture resolution and model complexity, processing time per image generally ranges from tens milliseconds to a second. Longer time may yield more exact findings, and vice versa. A few of proposal-based object detectors are initially slower but more accurate, though they may attain faster inference if the maximum number of region proposals is limited. This is not practical for ship identification problem, because harbor images are similar to the one occurred often. Real-time analysis is conducted based on the region of interest. As a result, a simple and non-complex model is frequently deemed for specific issues. A methodology like this would often result in a greater false alarm and miss rate if it is applied to ship detection.

The studied reference models include Faster R-CNN and YOLOv5. Faster R-CNN and YOLOv5 achieved the highest and lowest AP results, respectively. Model selection is frequently influenced by the speed of visual object recognition. Depending on image resolution and model complexity, processing time of every image generally ranges from ten milliseconds to a second. Longer time may yield more exact results, and vice versa. Visual object detectors related to proposals are initially slower but more precise, though they may attain faster detection if the maximum number of region proposals is limited.

We compared the approach to others, including Fast R-CNN, Faster R-CNN, YOLOv2, and SSD. We chose the VGG training detection model for the Fast R-CNN algorithm. Regarding Faster R-CNN, we made use of a pre-trained convolutional neural network on ImageNet as the pre-trained model, then we adopted ZF net (3 fully connected layers and 5 convolutional layers) as well as VGG-16 net (3 fully connected layers and 13 convolutional layers) to retrain the detection model.

In order to achieve the goal, we utilize VGG-16 net and MobileNet. Regarding YOLOv2, we retrain the detection model by using the pre-trained weights while increasing the quantity of data and improving model resilience with typical data augmentation methods like as hue, saturation, and exposure changes. Pertaining to model training, these parameters have been taken into account. All experiments were carried out based on Titan XP computers. We documented the outcomes of each model based on the prior indicators of evaluation. The proposed model relied on YOLOv5, the mAP of every model has been improved significantly.

The conventional object detection methods cannot work in real time. Fortunately, the emergence of deep learning has opened a new era of visual object detection. The experiments using basic deep learning methods show that the single-stage object detection method has significant advantages in visual object detection, the accuracy has made significant progress, which is conducive to improving the efficiency of object detection in real time.

One of the main aims of this thesis is to evaluate the appropriateness of algorithms of YOLO models in computer vision. The accuracy of YOLO is measured by comparing YOLO detections to observational methods. Human observation with video records determines the amount of passing boats. If the findings of human observation and machine observation were compared, it was discovered that there are several disparities in evaluations between these two approaches.

Human observation and YOLO models have the same assessment settings, which utilizes human observations as the baseline which is more logical. Due to human errors, a machine observation like YOLO models may potentially fail to visual object detection (Woods, Dekker, Cook, Johannesen, & Sarter, 2017). The detections were re-evaluated numerous times with various settings to reduce the mistakes of machine classification.

The results revealed that there is a significant disparity between human object detection and machine classification. Because human observation is a more exact source than machine classification, it will be utilized as the baseline (ground truth) for calculating the accuracy of YOLO. However, two sorts of errors are conceivable for calculating YOLO accuracy: Misclassification and misdetection, both are considered faults in the accuracy computations. YOLO models can accurately detect visual objects spotted by human observations. Visual object detection was carried out by using computer vision algorithms based on the distinguishing features from YOLO models (Kapur, 2017). Despite of various problems, sailboat classification has achieved a high accuracy by using YOLO models.

The effectiveness of ship detection by using YOLO models demonstrates its enormous potential in a variety of settings, including wind speeds, incident angles, ocean dynamic characteristics, and sea states, which primarily impact the backscattering coefficient between the ship and the ocean surface.

To get improved ship detection results, further effort will be required to identify the ships from images obtained from the satellites. In order to produce more precise detection results, the model can be updated, a new model can be sent to the satellite.

5.2.2 Validity and Reliability

In this thesis, the dependent variable is YOLO accuracy, the independent factors are the amount of training inputs. The validity approach is applied to examine the possibility of the ship detection methods. Counting the vessels and classifying them are possible by using the tools within the validity context.

In order to evaluate YOLO accuracy, it must be compared to benchmarks. The datasets such as PASCAL VOC (PASCAL, 2019) and COCO (COCO, 2019) are unsuitable for comparisons. These datasets were designed for general purposes of object detection, and no particular dataset for marine applications is provided. Furthermore, with these datasets, the accuracy of visual object detection and localization as a single metric is given, whereas the current research focus is only on visual object detection. As a result, a comparison of accuracies is needed.

Rodin et al. (2018) employed a CNN algorithm for visual object recognition during a marine search and rescue mission utilizing data gathered by an unmanned aerial vehicle and obtained 92.5% accuracy of visual object detection, as described in the literature. Furthermore, Yang et al. (2019) took use of YOLOv2 for visual object detection based on SAR photos for marine traffic surveillance and achieved a detection accuracy 90%. In this thesis, YOLO was employed for the detection task. As shown in Figure 4.6, Figure 4.7, Fgiure 4.8, it achieved an accuracy within the same range by utilizing a slightly different configuration.

5.2.3 Misclassification and Misdetection Errors

Misclassification and misdetection errors are the two types of YOLO detection faults. Misclassification occurs if YOLO assigns a vessel to the incorrect class, whereas misdetection occurs if YOLO is unable to identify a vessel. Both errors were treated as system faults in the accuracy computation. Assuming YOLO models were given sensory data to a collision avoidance system aboard an autonomous vessel, two degrees of repercussions are expected. In the instance of misdetection, the feedback is potentially more severe than in the case of misclassification. Misdetection of a rowboat with four paddlers, for example, has greater damaging consequences than misclassification of a rescue craft as a motorboat.

In this thesis, YOLO models obtained 94% accuracy in the test phase and 95% accuracy in the assessment phase. These results are regarded to be in the same ballpark as the benchmarks, but the inaccuracies must also be assessed from a safety standpoint. The mistakes occurred among those groups who are most vulnerable to severe effects in the event of an accident. This raises safety issues about the implementation.

Chapter 6
Conclusion and Future Work

In this chapter, we will summarize the various methods of this research project, analyze the shortcomings and gaps, and point out the directions of future improvement.

6.1 Conclusion

With the emergence of deep learning for evaluating visual object detection, the goal of this research project is to examine the available methods for ship detection. The key points of this thesis are as follows: We have reviewed image segmentation with specific applications, the related work has thoroughly reviewed with various applications, we presented our technical insights; we update deep learning methods for image segmentation; we review the work closely related to research gaps.

The first experiment in this thesis described the detection algorithm YOLO. The method is based on the automatic search, which has low costs on improving the detection accuracy and efficiency. Using sailboat images can distinguish them from other vessels, whether there are other vessels in this image.

The basic idea is to select our own feature vectors and optimize the attention of the model connection while maintaining the accurate prediction by using the regression algorithm, the whole network can better filter out its own vectors for subsequent regression. At the same time, in feature extraction networks, the observation mechanism and improved methods in NAS-SCAM-YOLO are transferred to the next step, this will be to visually explain the attention mechanism and more intuitively detail the overall characteristics of the models.

In the second project, we implement kayak and sailboat detection, we successfully mixed YOLOv5 and Transformer Backbone together for visual object detection. For each model, we find the best parameters such as input size, learning rate, and the best loss function. Finally, we ensemble these models and get a model with a few of cross-validation errors and generalization errors. The next stage of this project is to keep looking for other suitable backbone structures and attain a better result.

In this thesis, we discuss the result of YOLO parameters and successfully improve the YOLO performance by using ensembled models. In the next stage, we will keep improving the model performance by:

- Implementing more models and ensemble them together.
- The dataset has repetitive images, so cleaning dataset could also improve the

performance.

- Data augmentation will be further conducted, e.g., blurring, sharpening to increase the robust of the training set. It will make the unclearly images much better.
- After the data augmentation, TTA will also be conducted in the test set, which may improve the performance of ship detection.
- The general structure of visual object detection is a convolution backbone for image feature extraction, the output of the feedforward neural network is to generate the result of visual object detection. In this case, there may have other methods that could be applied to replace the attention method.

6.2 Future Work

Future work will be conducted to fill the existing research gap. Expanding the database will bring richer comparable data to this research, such as accuracy comparisons in scenarios with different levels of complexity. In addition, the trade-off between time and accuracy is also a valuable research topic in the field of navigation. A computational framework may be constructed to take into account of artifacts at this moment, which has never been done before. This will provide an opportunity to improve the preprocessing strategy and contribute to the process of visual feature extraction. Furthermore, a new deep learning model emphasizing on decreased training is needed to reduce the processing time.

References

Al-Sarayreh, M., Reis, M., Yan, W., Klette, R. (2019) A sequential CNN approach for foreign object detection in hyperspectral images. *International Conference on Information, Communications and Signal.*

Al-Sarayreha, M., Reis, M., Yan, W., Klette, R. (2020) Potential of deep learning and snapshot hyperspectral imaging for classification of species in meat. *Food Control*.

An, N., Yan, W. (2021) Multitarget tracking using Siamese neural networks. *ACM Transactions on Multimedia Computing, Communications and Applications.*

An, N. (2020) *Anomalies Detection and Tracking Using Siamese Neural Networks.* Master's Thesis. Auckland University of Technology, New Zealand.

Baker, B., Gupta, O., Raskar, R. (2016). Designing neural network architectures using reinforcement learning. arXiv:1611.02167.

Bhandare, A., Bhide, M., Gokhale, P., Chandavarkar, R. (2016). Applications of convolutional neural networks. International Journal of Computer Science and Information Technologies, Vol. 7 (5), 2206-2215.

Bi, F., Liu, F., Gao, L. (2010). A hierarchical salient-region based algorithm for ship detection in remote sensing images. Advances in Neural Network Research and Applications, Springer, pp. 729–738.

Bochkovskiy, A., Wang., C., and Liao, H., (2020). YOLOv4: Optimal speed and accuracy of object detection. arXiv:2004.10934v1.

Bouma, H., de Lange, J., van den Broek, S., Kemp, R.A., Schwering, P.B. (2008). Automatic detection of small surface targets with electro-optical sensors in a harbor environment. Electro-Optical Remote Sensing, Photonic Technologies, and Applications II, pp. 711402.

Cao, J., Chen, Q., Guo, J., and Shi, R., (2005). Attention-guided context feature pyramid network for object detection. arXiv: 2005.11475.

Cao, X. (2022) *Pose Estimation of Swimmers from Digital Images Using Deep Learning*. Master's Thesis, Auckland University of Technology.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S. (2020). End-to-end object detection with Transformers. arXiv.2005.12872.

Chang, Y.-L., Anagaw, A., Chang, L., Wang, Y., Hsiao, C.-Y. and Lee, W.-H. (2019). Ship detection based on YOLOv2 for SAR imagery. Remote Sensing, 11(7), p.786.

Chen, L., Shi, W., & Deng, D. (2021). Improved YOLOv3 based on attention mechanism for fast and accurate ship detection in optical remote sensing images. Remote Sensing, 13(4), 660.

Chen, S., Tao, C., Wang, X., Xiao, S. (2018). Polarimetric SAR targets detection and classification with deep convolutional neural network. PIERS-Toyama, pp. 2227–2234.

Chen, W., Li, J., Xing, J., Yang, Q., Zhou, Q. A. (2018). maritime targets detection method based on hierarchical and multi-scale deep convolutional neural network. International Conference on Digital Image Processing, pp. 1080616.

Chen, Z., Li, B., Tian, L.F., Chao, D. (2017). Automatic detection and tracking of ship based on mean shift in corrected video sequences. International Conference on Image, Vision and Computing, China, pp. 449–453.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv:1412.3555.

Cireşan, C., Giusti, A., Gambardella, M., Schmidhuber, J. (2013). Mitosis detection in breast cancer histology images with deep neural networks BT. Medical Image Computing and Computer-Assisted Intervention, pp. 411–418.

Cireşan, D., Meier, U., Masci, J., Schmidhuber, J. (2012). Multi-column deep neural network for traffic sign classification. Neural Networks, 32:333–338.

Cui, W., Yan, W. (2016) A scheme for face recognition in complex environments. *International Journal of Digital Crime and Forensics* (IJDCF) 8 (1), 26-36.

Cui, W. (2015) *A Scheme of Human Face Recognition in Complex Environments*. Master's Thesis, Auckland University of Technology.

Cummins, F., Gers. F, A., Schmidhuber, J. (1999). Language identification from prosody without explicit features. EUROSPEECH'99, volume 1, pp. 371–374.

Dai, J., Li, Y., He, K., Sun, J. (2016). RFCN: Object detection via region based fully convolutional networks. NIPS (pp. 379–387).

Dai, J., Li, Y., He, K., Sun, J. (2016). R-FCN: object detection via region-based fully convolutional networks. Int. Conf. Neural Inf. Process. Syst., pp. 379–387, Barcelona, Spain.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., & Li, F. (2009). ImageNet: A large scale hierarchical image database. CVPR, pp. 248–255.

Deng, J., Lu, Y., Lee, V. (2021). Imaging-based crack detection on concrete surfaces using YOLO network. Structural Health Monitoring, 20(2), 484-499.

Deng, L., Yu, D., Delft, B. (2013). Deep learning: Methods and applications foundations and trends R in signal processing. Signal Processing, 7:3–4.

Devlin, J., Chang, M., Lee, K., Toutanova, K. (2018). BERT: Pre-training of deep bidirectional Transformers for language understanding. arXiv:1810.04805.

Devlin, J., Chang, W., Lee, K., Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. The Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171–4186.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2021). An image is worth 16 × 16 words: Transformers for image recognition at scale. arXiv.2010.11929.

Du, Y., Sun, S., Qiu, S., Li, S., Pan, M., Chen, C-H. (2021). Intelligent recognition system based on contour accentuation for navigation marks. Wireless Communications and Mobile Computing, pp. 1–11.

El-Darymli, K., McGuire, P., Power, D., Moloney, C.R. (2013). Target detection in synthetic aperture radar imagery: A state-of-the-art survey. J. Appl. Remote Sens, 7, pp. 7–35.

El-Darymli, K., Gill, E.W., Mcguire, P., Power, D., Moloney, C. (2016). Automatic target recognition in synthetic aperture radar imagery: A state-of-the-art review. IEEE Access, 4, 6014–6058.

Erhan, D., Szegedy, C., Toshev, A., & Anguelov, D. (2014). Scalable object detection using deep neural networks. CVPR (pp. 2147–2154).

Everingham, M., Eslami, S., Gool, L. V., Williams, C., Winn, J., & Zisserman, A. (2015). The pascal visual object classes challenge: A retrospective. IJCV, 111(1), 98–136.

Fefilatyev, S., Goldgof, D., Shreve, M., Lembke, C. (2012). Detection and tracking of ships in open sea with rapidly moving buoy-mounted camera system. Ocean Eng., 54, 1–12.

Felzenszwalb, P., Girshick, R., McAllester, D., & Ramanan, D. (2010). Object detection with discriminatively trained part-based models. IEEE TPAMI, 32(9), 1627–1645.

Fu, C.-Y., Liu, W., Ranga, A., Tyagi, A., Berg, A.C. (2013). DSDD: Deconvolutional single shot detector. arXiv:1701.06659.

Fu, Y., Nguyen, M., Yan, W. (2022) Grading methods for fruit freshness based on deep learning. *Springer Nature Computer Science.*

Fu, Y. (2020) *Fruit Freshness Grading Using Deep Learning.* Master's Thesis. Auckland University of Technology, New Zealand.

Gao, X., Nguyen, M., Yan, W. (2021) Face image inpainting based on generative adversarial network. *International Conference on Image and Vision Computing*

New Zealand.

Gao, X., Nguyen, M., Yan, W. (2022) A face image inpainting method based on autoencoder and adversarial generative networks. Pacific-Rim Symposium on Image and Video Technology.

Garc´ıa-Dom´ınguez, A. (2015). Mobile applications, cloud and bigdata on ships and shore stations for increased safety on marine traffic: A smart ship project. International Conference on Industrial Technology, pp. 1532–1537, Spain.

Gers, F. A., Schraudolph, N. N., Schmidhuber, J. (2002). Learning precise timing with LSTM recurrent networks. Journal of Machine Learning Research, 3, 115-143.

Ghiasi, G., Lin. T, Y., and Le. Q, V., (2019). NAS-FPN: Learning scalable feature pyramid architecture for object detection. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7036-7045.

Girshick, R. (2015). Fast R-CNN. IEEE International Conference on Computer Vision, pp. 1440–1448.

Girshick, R., Donahue, J., Darrell, T., Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587.

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2016). Region-based convolutional networks for accurate object detection and segmentation. IEEE TPAMI, 38(1), 142–158.

Girshick, R., Iandola, F., Darrell, T., & Malik, J. (2015). Deformable part models are convolutional neural networks. CVPR (pp. 437–446).

Gowdra, N. (2021) *Entropy-Based Optimization Strategies for Convolutional Neural Networks.* PhD Thesis, Auckland University of Technology, New Zealand.

Gu, D., Nguyen, M., Yan, W. (2016) Cross models for twin recognition. *International Journal of Digital Crime and Forensics* 8 (4), 26-36.

Gu, Q., Yang, J., Kong, L., Yan, W., Klette, R. (2017) Embedded and real-time vehicle detection system for challenging on-road scenes. *Optical Engineering*, 56 (6), 063102.

Gu, Q., Yang, J., Yan, W., Klette, R. (2017) Integrated multi-scale event verification in an augmented foreground motion space. Pacific-Rim Symposium on Image and Video Technology (pp.488-500)

Gu, Q., Yang, J., Yan, W., Li, Y., Klette, R. (2017) Local Fast R-CNN flow for object-centric event recognition in complex traffic scenes. Pacific-Rim Symposium on Image and Video Technology (pp.439-452)

Hariharan, B., Arbeláez, P., Girshick, R., Malik, J. (2016). Object instance segmentation and fine-grained localization using hypercolumns. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(4), 627–639.

Ho, Y. and Wookey, S. (2020). The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling. IEEE Access, 8, pp.4806–4813.

He, K., Zhang, X., Ren, S., & Sun, J. (2014). Spatial pyramid pooling in deep convolutional networks for visual recognition. ECCV (pp. 346–361).

Hoiem, D., Chodpathumwan, Y., & Dai, Q. (2012). Diagnosing error in object detectors. ECCV, pp. 340–353.

Hu, Z., Qin, H., Peng, X., Yue, T., Yue, H., Luo, G., Zhu, W. (2019). Infrared polymorphic target recognition based on single step cascade neural network. AI in Optics and Photonics, pp. 113420T.

Huang, J., Wang, J., Tan, Y., Wu, D., and Cao, Y. (2020). An automatic analog instrument reading system using computer vision and inspection robot. IEEE Transactions on Instrumentation and Measurement, vol. 69, no. 9, pp. 6322–6335.

Im, I., Shin, D., & Jeong, J. (2018). Components for smart autonomous ship architecture based on intelligent information technology. Procedia Computer Science, 134, 91–98.

Jeong, C., Yang, H.S., Moon, K. (2018). Horizon detection in maritime images using scene parsing network. Electron. Lett., 54, 760–762.

Jiao, J., Zhang, Y., Sun, H., Yang, X., Gao, X., Hong, W., Fu, K. and Sun, X. (2018). A densely connected end-to-end neural network for multiscale and multiscene SAR ship detection. IEEE Access, 6, pp. 20881–20892.

Jiao, L., Zhang, F., Liu, F., Yang, S., Li, L., Feng, Z., Qu, R. (2019). A survey of deep learning-based object detection. IEEE Access, 7, 128837–128868.

Jiao, Y., Weir, J., Yan, W. (2011) Flame detection in surveillance. Journal of Multimedia 6 (1).

Jie, H., Li, S., Gang, S., and Albanie, S., (2017). Squeeze-and-excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, pp.99.

Jie, Y., Leonidas, L., Mumtaz, F. and Ali, M. (2021). Ship detection and tracking in inland waterways using improved YOLOv3 and deep SORT. Symmetry, 13(2), pp.308.

Kang, M., Leng, X., Lin, Z., Ji, K. (2017). A modified Faster R-CNN based on CFAR algorithm for SAR ship detection. International Workshop on Remote Sensing with Intelligent Processing, pp. 1–4.

Kanjir, U., Greidanus, H., Oštir, K. (2018). Vessel detection and classification from spaceborne optical images: A literature survey. Remote Sens. Environ, 207, 1–26.

Khan, A., Zahoora, A., Qureshi, A. (2017). A survey of the recent architectures of deep convolutional neural networks. Artificial Intelligence Review. 53(8), 5455-5516.

Krizhevsky, A., and Hinton, G. (2009). Learning multiple layers of features from tiny images. Handbook of Systemic Autoimmune Diseases, 1 (4).

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., PontTuset, J., et al. (2020). The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. International Journal of Computer Vision, 128, pp. 1956–1981.

Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. IEEE CVPR, pp. 2169–2178.

Le, R., Nguyen, M., Yan, W. (2020) Machine learning with synthetic data – a new way to learn and classify the pictorial augmented reality markers in real-time. International Conference on Image and Vision Computing New Zealand.

Le, R., Nguyen, M., Yan, W. (2021) Training a convolutional neural network for transportation sign detection using synthetic dataset. *International Conference on Image and Vision Computing New Zealand.*

Le, R., Nguyen, M., Yan, W., Nguyen, H. (2021) Augmented reality and machine learning incorporation using YOLOv3 and ARKit. *Applied Sciences.*

Le, R. (2022) *Synthetic Data Annotation for Enhancing the Experiences of Augmented Reality Application Based on Machine Learning (PhD Thesis).* Auckland University of Technology, New Zealand.

Li, C., Yan, W. (2021) Braille recognition using deep learning. *International Conference on Control and Computer Vision.*

Li. C. (2022) *Special Character Recognition Using Deep Learning. Master's Thesis* Auckland University of Technology, New Zealand.

Li, P. (2018) *Rotation Correction for License Plate Recognition*. Master's Thesis, Auckland University of Technology, New Zealand.

Li, P., Nguyen, M., Yan, W. (2018) Rotation correction for license plate recognition. *International Conference on Control, Automation and Robotics*.

Li, R., Nguyen, M., Yan, W. (2017) Morse codes enter using finger gesture recognition. *International Conference on Digital Image Computing: Techniques and Applications*.

Li, Y., Li, Z., Zhu, Y., Li, B., Xiong, W., Huang, Y. (2019). Thermal infrared small ship detection in sea clutter based on morphological reconstruction and multi-feature analysis. Appl. Sci., 9, 3786.

Lin, C., Chen, W., Zhou, H. (2020). Multi-visual feature saliency detection for sea-surface targets through improved sea-skyline detection. Journal of Marine Science and Engineering, 8(10), 799.

Lin, J., Yu, Q., Chen, G. (2019). Infrared ship target detection based on the combination of Bayesian theory and SVM. Automatic Target Recognition and Navigation, (Vol. 11429, pp. 1142919).

Lin, T., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, L. (2014). Microsoft COCO: Common objects in context. ECCV (pp. 740–755).

Lin, T.-Y., Goyal, P., Girshick, R., He, K. and Dollar, P. (2018). Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, pp.1–1.

Lin, Z., Ji, K., Leng, X., Kuang, G. (2019). Squeeze and excitation rank Faster R-CNN for ship detection in SAR images. IEEE Geosci. Remote Sens. Lett., 16, 751–755.

Lipschutz, I., Gershikov, E., Milgrom, B. (2013). New methods for horizon line detection in infrared and visible sea images. Int. J. Comput. Eng. Res, 3, 1197–1215.

Litjens, G., Kooi, T., Bejnordi, B., Setio, A., Ciompi, F., Ghafoorian, M., et al. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 60–88.

Liu, M., Yan, W. (2022) Masked face recognition in real-time using MobileNetV2. *ACM ICCCV.*

Liu, X., Nguyen, M., Yan, W. (2019) Vehicle-related scene understanding using deep learning. *Asian Conference on Pattern Recognition.*

Liu, X. (2019) *Vehicle-related Scene Understanding Using Deep Learning.* Master's Thesis, Auckland University of Technology, New Zealand.

Liu, X., Yan, W. (2020) Vehicle-related scene segmentation using CapsNets. *International Conference on Image and Vision Computing New Zealand.*

Liu, X., Yan, W. (2021) Traffic-light sign recognition using Capsule network. *Springer Multimedia Tools and Applications.*

Liu, Z., Yan, W., Yang, B. (2018) Image denoising based on a CNN model. International Conference on Control, Automation and Robotics.

Liu, H., Hou, X. (2012). Moving detection research of background frame difference based on Gaussian model. International Conference on Computer Science & Service System, (pp. 258-261).

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Fu, C-Y., and Berg, A. C. (2016). SSD: single shot multibox detector. Proc. Eur. Conf. Comput. Vis., pp. 21–37, The Netherlands.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C. (2016). SSD: Single shot multibox detector. European Conference on Computer Vision, pp. 21–37.

Liu, Z., Wang, H., Zhang, S., Wang, G., and Qi, J. (2020). NAS-SCAM: Neural architecture search-based spatial and channel joint attention module for nuclei semantic segmentation and classification. MICCAI, pp. 263-272.

Liu, Z., Jiang, T., Zhang, T., Li, Y. (2019). IR ship target saliency detection based on lightweight non-local depth features. International Conference on Electronic Information Technology and Computer Engineering, pp. 1681–1686.

Lu, J., He, Y., Li, H.-Y., Lu, F. (2006). Detecting small target of ship at sea by infrared image. IEEE International Conference on Automation Science and Engineering, pp. 165–169.

Lu, J. (2016) Empirical Approaches for Human Behavior Analytics. Master's Thesis. Auckland University of Technology, New Zealand.

Lu, J., Nguyen, M., Yan, W. (2018) Pedestrian detection using deep learning. IEEE AVSS.

Lu, J., Nguyen, M., Yan, W. (2020) Human behavior recognition using deep learning. *International Conference on Image and Vision Computing New Zealand*.

Lu, J., Nguyen, M., Yan, W. (2021) Sign language recognition from digital videos using deep learning methods. *International Symposium on Geometry and Vision.*

Lu, J. (2021) *Deep Learning Methods for Human Behavior Recognition*. PhD Thesis. Auckland University of Technology, New Zealand.

Luo, Z., Nguyen, M., Yan, W. (2022) Kayak and sailboat detection based on the improved YOLO with Transformer. *ACM ICCCV.*

Luo, Z., Nguyen, M., Yan, W. Sailboat detection based on automated search attention mechanism and deep learning models. *International Conference on Image and Vision Computing New Zealand.*

Ma, X. (2020) *Banknote Serial Number Recognition Using Deep Learning.* Master's Thesis, Auckland University of Technology, New Zealand.

Ma, X., Yan, W. (2021) Banknote serial number recognition using deep learning. *Springer Multimedia Tools and Applications*.

Mehtab, S., Yan, W. (2021) FlexiNet: Fast and accurate vehicle detection for autonomous vehicles-2D vehicle detection using deep neural network. *International*

Conference on Control and Computer Vision.

Mehtab, S., Yan, W. (2022) Flexible neural network for fast and accurate road scene perception. *Multimedia Tools and Applications.*

Mehtab, S. Yan, W., Narayanan, A. (2022) 3D vehicle detection using cheap LiDAR and camera sensors. *International Conference on Image and Vision Computing New Zealand.*

Marino, A. (2013). A notch filter for ship detection with polarimetric SAR data. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., 6, 1219–1232.

Moreira, A., Prats-Iraola, P., Younis, M., Krieger, G., Hajnsek, I., Papathanassiou, K.P. (2013). A tutorial on synthetic aperture radar. IEEE Geosci. Remote Sens. Mag., 1, 6–43.

Narendra, K. S., & Parthasarathy, K. (1990). Identification and control of dynamical systems using neural networks. IEEE Transactions on Neural Networks, 1(1), 4-27.

Nguyen, M., Yan, W. Temporal colour-coded facial-expression recognition using convolutional neural network. *International Summit Smart City 360°: Science and Technologies for Smart Cities.*

Niitsuma, M., Tomita, Y., Yan, W., Bell, D. (2018) Towards musicologist-driven mining of handwritten scores. *IEEE Intelligent Systems*.

Niitsuma, M., Tomita, Y., Yan, W., Bell, D. (2011) Classifying Bach's handwritten C-Clefs. *International Society for Music Information Retrieval Conference (ISMIR)*.

Oksuz, K., Cam, B., Kalkan, S., Akbas, E. (2020). Imbalance problems in object detection: A review. IEEE Trans. Pattern Anal. Mach. Intell., 43, 3388–3415.

Oliver, C., Quegan, S. (2004). Understanding synthetic aperture radar images. SciTech Publishing. pp. 427-429.

Özertem, K.A. (2016). A fast automatic target detection method for detecting ships in infrared scenes. Proceedings of the Automatic Target Recognition XXVI, pp. 984404.

Pan, C., Yan, W. (2018) A learning-based positive feedback in salient object detection. International Conference on Image and Vision Computing New Zealand.

Pan, C., Yan, W. (2020) Object detection based on saturation of visual perception. *Multimedia Tools and Applications*, 79 (27-28), 19925-19944.

Pan, C., Liu, J., Yan, W., Zhou, Y. (2021) Salient object detection based on visual perceptual saturation and two-stream hybrid networks. *IEEE Transactions on Image Processing.*

Pan, M., Liu, Y., Cao, J., Li, Y., Li, C. and Chen, C.-H. (2020). Visual recognition based on deep learning for navigation mark classification. IEEE Access, 8, pp.32767–32775.

Park, J., Woo, S., Lee, J. Y., and Kweon, I. S. (2018). BAM: Bottleneck attention module, arXiv: 1807.06514.

Pham, H., Guan, M. Y., Zoph, B., Le, Q. Y., and Dean, J., (2018). Efficient neural architecture search via parameter sharing. International Conference on Machine Learning, pp. 4095-4104.

Prasad, D.K., Rajan, D., Rachmawati, L., Rajabally, E., Quek, C. (2017). Video processing from electro-optical sensors for object detection and tracking in a maritime environment: A survey. IEEE Trans. Intell. Transp. Syst., 18, 1993–2016.

Qi, J., Nguyen, M., Yan, W. (2022) Waste classification from digital images using ConvNeXt. Pacific-Rim Symposium on Image and Video Technology.

Qi, L., Chen, W., Dong, J., Huang, G., Xue, W. (2019). Ship target detection algorithm based on improved Faster R-CNN. Electronics, 8(9), p.959.

Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O. R., and Jagersand, M. (2020). U2-Net: Going deeper with nested U-structure for salient object detection. Pattern Recognition, vol. 106, pp. 107404.

Qin, Z., Yan, W. (2021) Traffic-sign recognition using deep learning. *International Symposium on Geometry and Vision.*

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788.

Redmon, J., Farhadi, A. (2017). YOLO9000: Better, faster, stronger. IEEE Conference on Computer Vision and Pattern Recognition, pp.7263–7271.

Redmon, J.; Farhadi, A. (2018). YOLOv3: An incremental improvement. arXiv:1804.02767.

Ren, S., He, K., Girshick, R. and Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6), pp.1137–1149.

Ren, Y., Nguyen, M., Yan, W. (2018) Real-time recognition of series seven New Zealand banknotes. *International Journal of Digital Crime and Forensics* (IJDCF) 10 (3), 50-66.

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., and Savarese, S. (2019). Generalized intersection over union: A metric and a loss for bounding box regression.

IEEE/CVF Conference on Computer Vision and Pattern Recognition. (pp. 658-666).

Roy. A, G., Nav Ab, N., and Wachinger, C. (2018). Concurrent spatial and channel squeeze & excitation in fully convolutional networks. International Conference on Medical Image Computing and Computer-Assisted Intervention. (pp. 421-429). Springer.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). ImageNet large scale visual recognition challenge. IJCV, 115(3), 211–252.

Sande, V., Uijlings, J. R., Gevers, T., and Smeulders, A. W. (2011). Segmentation as selective search for object recognition. IEEE International Conference on Computer Vision (ICCV), pp. 1879–1886.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., & LeCun, Y. (2014). OverFeat: Integrated recognition, localization and detection using convolutional networks. arXiv:1312.6229.

Shadman, S., Nazib A., Ahmed J., Hridon, A. (2018). An overview of convolutional neural network: Its architecture and applications. Preprints 2018110546.

Shan, X., Zhao, D., Pan, M., Wang, D., Zhao, L. (2019). Sea-skyline and its nearby ships detection based on the motion attitude of visible light sensors. Sensors, 19, 4004.

Shan, Y., Zhou, X., Liu, S., Zhang, Y., Huang, K. (2020). SiamFPN: A deep learning method for accurate and real-time maritime ship tracking. IEEE Trans. Circuits Syst. Video Technol., 31, 315–325.

Shao, Z., Wu, W., Wang, Z., Du, W., Li, C., (2018). SeaShips: A large-scale precisely annotated dataset for ship detection. IEEE Transactions on Multimedia, 20 (10), 2593-2604.

Shen, D., Xin, C., Nguyen, M., Yan, W. (2018) Flame detection using deep learning. *International Conference on Control, Automation and Robotics.*

Shen, H., Kankanhalli, M., Srinivasan, S., Yan, W. (2004) Mosaic-based view enlargement for moving objects in motion pictures. *IEEE ICME'04*.

Shen, Y., Yan, W. (2019) Blind spot monitoring using deep learning. *International Conference on Image and Vision Computing New Zealand*.

Song, C., He, L., Yan, W., Nand, P. (2019) An improved selective facial extraction model for age estimation. *International Conference on Image and Vision Computing New Zealand*.

Simonyan, K., Zisserman, A. (2015). Very deep convolutional networks for large scale image recognition. arXiv:1409.1556.

Singh, R., Vashisht, M., Qamar, S. (2017). Role of linguistic quantifier and digitally approximated Laplace operator in infrared based ship detection. Int. J. Syst. Assur. Eng. Manag., 8, 1336–1342.

Smith, M., Varshney, P. (2000). Intelligent CFAR processor based on data variability. IEEE Trans. Aerosp. Electron. Syst., 36, 837–847.

Spanhol, F., Oliveira, L., Petitjean, C., Heutte, L. (2016). A dataset for breast cancer histopathological image classification. IEEE Trans Biomed Eng, 63:1455–1462.

Sun, Y.-Q., Tian, J.-W., Liu, J. (2005). Background suppression based-on wavelet transformation to detect infrared target. International Conference on Machine Learning and Cybernetics, pp. 4611–4615.

Sutskever, I., Vinyals, O., Le, Q. (2014). Sequence to sequence learning with neural networks. Advances in Neural Information Processing Systems, pp 3104–3112.

Szegedy, C., Toshev, A., & Erhan, D. (2013). Deep neural networks for object detection. NIPS (pp. 2553–2561).

Tang, D., Sun, G., Wang, D.-H., Niu, Z.-D., Chen, Z.-P. (2013). Research on infrared ship detection method in sea-sky background. International Symposium on Photoelectronic Detection and Imaging: Infrared Imaging and Applications, pp. 89072H.

Tao, D., Doulgeris, A., Brekke, C. (2016). A segmentation-based CFAR detection algorithm using truncated statistics. IEEE Trans. Geosci. Remote Sens., 54, 2887–2898.

Tong, D., Yan, W. (2022) Visual watermark identification from the transparent window of currency by using deep learning. *Applications of Encryption and Watermarking for Information Security.*

Uijlings, J., van de Sande, K., Gevers, T., & Smeulders, A. (2013). Selective search for object recognition. IJCV, 104(2), 154–171.

Vachon, P. W., Campbell, J. W. M., Bjerkelund, C. A., Dobson, F. W., & Rey, M. T. (1997). Ship detection by the RADARSAT SAR: Validation of detection model predictions. Canadian Journal of Remote Sensing, 23(1), 48–59.

Vedaldi, A., Gulshan, V., Varma, M., & Zisserman, A. (2009). Multiple kernels for object detection. ICCV (pp. 606–613).

Varnima, E., Ramachandran, C. (2020). Real-time Gender Identification from face images using YOLO. International Conference on Trends in Electronics and Informatics (ICOEI), pp. 1074-1077

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł., Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, pp. 5998–6008.

Wang, A., Wang, C., Su, W., Dong, Y. (2010). Adaptive segmentation algorithm for ship target under complex background. International Conference on Advanced Computer Theory and Engineering, pp. V2-219–V2-223.

Wang, G., Wu, X., Yan, W. (2017) The state-of-the-art technology of currency identification: A comparative study. International Journal of Digital Crime and Forensics 9 (3), 58-72.

Wang, H., Yan, W. (2022) Face detection and recognition from distance based on deep learning. *Aiding Forensic Investigation Through Deep Learning and Machine Learning Framework.* IGI Global.

Wang, J., Yan, W., Kankanhalli, M., Jain, R., Reinders, M. (2003) Adaptive monitoring for video surveillance. International Conference on Information, Communications and Signal Processing.

Wang, J., Kankanhalli, M., Yan, W., Jain, R. (2003) Experiential sampling for video surveillance. *ACM SIGMM International Workshop on Video surveillance* (pp.77-86).

Wang, J., Yan, W. (2016) BP-neural network for plate number recognition. International *Journal of Digital Crime and Forensics* (IJDCF) 8 (3), 34-45.

Wang, J. (2016) *Event-driven Traffic Ticketing System*. Master's Thesis, Auckland University of Technology, New Zealand.

Wang, J., Bacic, B., Yan, W. (2018) An effective method for plate number recognition. *Multimedia Tools and Applications,* 77 (2), 1679-1692.

Wang, L., Yan, W. (2021) Tree leaves detection based on deep learning. *International Symposium on Geometry and Vision.*

Wang, X., Yan, W. (2019) Cross-view gait recognition through ensemble learning. *Neural Computing and Applications.*

Wang, X., Yan, W. (2019) Gait recognition using multichannel convolutional neural networks. *Neural Computing and Applications.*

Wang, X., Yan, W. (2019) Multi-perspective gait recognition based on ensemble learning. *Springer Neural Computing and Applications*.

Wang, X., Yan, W. (2019) Human gait recognition based on frame-by-frame gait energy images and convolutional long short-term memory. *International Journal of Neural Systems.*

Wang, X., Yan, W. (2020) Non-local gait feature extraction and human identification.

Springer Multimedia Tools and Applications.

Wang, X., Yan, W. (2020) Cross-view gait recognition through ensemble learning. *Neural computing and applications* 32 (11), 7275-7287.

Wang, X., Yan, W. (2022) Human identification based on gait manifold. *Applied Intelligence.*

Wang, Y., Wang, C., Zhang, H., Dong, Y. and Wei, S. (2019). A SAR dataset of ship detection for deep learning under complex backgrounds. Remote Sensing, 11(7), p.765.

Wang, Y., Wang, C., Zhang, H., Dong, Y., Wei, S. (2019). A SAR dataset of ship detection for deep learning under complex backgrounds. Remote Sens., 11, 765.

Wawrzyniak, N., Hyla, T. and Popik, A. (2019). Vessel detection and tracking method based on video surveillance. Sensors, 19(23), p.5230.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., Yu, P. S. (2019). A comprehensive survey on graph neural networks. IEEE Transactions on Neural Networks and Learning Systems, 32(1), 4-24.

Xiang, Y., Yan, W. (2021) Fast-moving coin recognition using deep learning. *Springer Multimedia Tools and Applications.*

Xiao, B., Nguyen, M., Yan, W. (2021) Apple ripeness identification using deep learning. *International Symposium on Geometry and Vision.*

Xiao, B., Nguyen, M., Yan, W. (2022) *Fruit ripeness identification using Transformer model SSRN.*

Xin, C. (2020) *Detection and Recognition for Multiple Flames Using Deep Learning.* Master's Auckland University of Technology, New Zealand.

Xin, C., Nguyen, M., Yan, W. (2020) Multiple flames recognition using deep learning. *Handbook of Research on Multimedia Cyber Security*, 296-307.

Xing, J., Yan, W. (2021) Traffic sign recognition using guided image filtering. *International Symposium on Geometry and Vision.*

Xing, J., Nguyen, M., Yan, W. (2022) The improved framework of traffic sign recognition by using guided image filtering. *Springer Nature Computer Science.*

Xing, J. (2022) *Traffic Sign Recognition from Digital Images Using Deep Learning.* Master's Thesis, Auckland University of Technology, New Zealand.

Xing, J., Nguyen, M., Yan, W. (2022) Traffic sign recognition from digital images by using deep learning. Pacific-Rim Symposium on Image and Video Technology.

Xu, G., Wang, J., Qi, S. (2017). Ship detection based on rotation-invariant HOG descriptors for airborne infrared images. Pattern Recognition and Computer Vision, pp. 1060912.

Xu, H., Yao, L., Zhang, W., Liang, X., and Li, Z. (2019). Auto-FPN: Automatic network architecture adaptation for object detection beyond classification. IEEE/CVF International Conference on Computer Vision, pp. 6649-6658.

Xu, Y., Yu, G., Wang, Y., Wu, X. and Ma, Y. (2017). Car detection from low-altitude UAV imagery with the Faster R-CNN. Journal of Advanced Transportation, pp.1–10.

Yan, W., Kankanhalli, M. (2002) Detection and removal of lighting & shaking artifacts in home videos. *ACM International Conference on Multimedia*, 107-116.

Yan, W., Wang, J., Kankanhalli, M. (2005) Automatic video logo detection and removal. *Multimedia Systems* 10 (5), 379-391.

Yan, W., Chambers, J. (2013) An empirical approach for digital currency forensics. *IEEE International Symposium on Circuits and Systems* (ISCAS), 2988-2991.

Yan, W., Chambers, J., Garhwal, A. (2014) An empirical approach for currency identification. *Multimedia Tools and Applications* 74 (7).

Yan, W., Kankanhalli, M. (2015) Face search in encrypted domain. Pacific-Rim Symposium on Image and Video Technology, 775-790.

Yan, W. (2019) *Introduction to Intelligent Surveillance: Surveillance Data Capture, Transmission, and Analytics*. Springer London.

Yan, W. (2021) *Computational Methods for Deep Learning: Theoretic, Practice and Applications*. Springer London.

Yang, W., Li, H., Liu, J., Xie, S., Luo, J. (2019). A sea-sky-line detection method based on Gaussian mixture models and image texture features. Int. J. Adv. Robot. Syst., 16, 1–12.

Young, T., Hazarika, D., Poria, S., Cambria, E. (2018). Recent trends in deep learning based natural language processing. IEEE Computational Intelligence, 13(3), 55–75.

Yu, Z. (2021) *Deep Learning Methods for Human Action Recognition*. Master's Thesis, Auckland University of Technology, New Zealand.

Yu, Z., Yan, W. (2020) Human action recognition using deep learning methods. *International Conference on Image and Vision Computing New Zealand.*

Zhang, J., Wu, J., Wang, H., Wang, Y. and Li, Y. (2021). Cloud detection method using

CNN based on cascaded feature attention and channel attention. IEEE Transactions on Geoscience and Remote Sensing, pp.1–1.

Zhang, L., Lin, W. (2013). Background of visual attention - theory and experiments. IEEE Selective Visual Attention: Computational Models and Applications, pp.25-71.

Zhang, L. (2020) *Virus Identification from Digital Images Using Deep Learning*. Master's Thesis, Auckland University of Technology, New Zealand.

Zhang, L., Yan, W. (2020) Deep learning methods for virus identification from digital images. *International Conference on Image and Vision Computing New Zealand*.

Zhang, Q. (2018) *Currency Recognition Using Deep Learning*. Master's Thesis, Auckland University of Technology, New Zealand.

Zhang, Q., Yan, W. (2018) Currency detection and recognition based on deep learning. *IEEE International Conference on Advanced Video and Signal Based Surveillance*.

Zhang, Q., Yan, W., Kankanhalli, M. (2019) Overview of currency recognition using deep learning. *Journal of Banking and Financial Technology,* 3 (1), 59–69.

Zhang, S., Wu, R., Xu, K., Wang, J., Sun, W. (2019). R-CNN-based ship detection from high resolution remote sensing imagery. Remote Sens., 11, 631.

Zhang, X., Yang, Y., Han, Z., Wang, H., & Gao, C. (2013). Object class detection: A survey. ACM Computing Surveys, 46(1), 10:1–10:53.

Zhang, Y., Yan, W., Narayanan, A. (2017) A virtual keyboard implementation using finger recognition. *International Conference on Image and Vision Computing New Zealand.*

Zhang, Y. (2016) *A Virtual Keyboard Implementation Based on Finger Recognition.* Master's Thesis, Auckland University of Technology, New Zealand.

Zhang, Z., Geiger, J., Pohjalainen, J., Mousa, A. E., Jin, W., Schuller, B. (2018). Deep learning for environmentally robust speech recognition: An overview of recent developments. ACM Transactions on Intelligent Systems and Technology, 9(5), 49:1–49:28.

Zhao, J., Zhang, Z., Yu, W. and Truong, T.-K. (2018). A cascade coupled convolutional neural network guided visual attention method for ship detection from SAR images. IEEE Access, 6, pp.50693–50708.

Zhao, K. (2021) *Fruit Detection Using CenterNet*. Master's Thesis, Auckland University of Technology, New Zealand.

Zhao, K., Yan, W. (2021) Fruit detection from digital images using CenterNet. *International Symposium on Geometry and Vision.*

Zhao, Y., Zhao, L., Xiong, B., Kuang, G. (2020). Attention receptive pyramid network for ship detection in SAR images. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., 13, 2738–2756.

Zheng, K., Yan, Q., Nand, P. (2017) Video dynamics detection using deep neural networks. *IEEE Transactions on Emerging Topics in Computational Intelligence*.

Zhou, A., Xie, W., Pei, J. (2018). Infrared maritime target detection using the high order statistic filtering in fractional Fourier domain. Infrared Phys. Technol, 91, 123–136.

Zhou, J., Cui, G., Zhang, Z., Yang, C., Liu, Z., Sun, M. (2018). Graph neural networks: A review of methods and applications. AI Open, 1, 57-81.

Zhou, Y., Tao, C. (2020). Multi-task BERT for problem difficulty prediction. International Conference on Communications, Information System and Computer Engineering (CISCE), pp. 213-216.

Zhu, X., Tuia, D., Mou, L., Xia, G., Zhang, L., Xu, F., et al. (2017). Deep learning in remote sensing: A comprehensive review and list of resources. IEEE Geoscience and Remote Sensing Magazine, 5(4), 8–36.

Zhu, Y., Yan, W. (2022) Ski fall detection from digital images using deep learning. *ACM ICCCV.*

Zhu, Y., Yan, W. (2022) Image-based storytelling using deep learning. *ACM ICCCV*.

Zhu, Y., Yan, W. (2022) Parasite detection from digital images using deep learning methods. Machine Learning and AI Techniques in Interactive Medical Image Analysis, IGI Global.

Zhu, Y., Yan, W. (2022) Traffic sign recognition based on deep learning. *Multimedia Tools and Applications.*